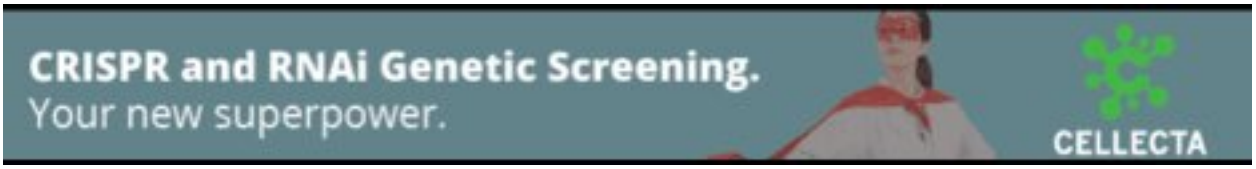
Rapid evolution of piRNA clusters in the *Drosophila melanogaster* ovary

Satyam Srivastav, Cedric Feschotte and Andrew Clark

Genome Res. published online May 15, 2024
Access the most recent version at doi:10.1101/gr.278062.123



Rapid evolution of piRNA clusters in the *Drosophila melanogaster* ovary

Satyam Srivastav*, Cédric Feschotte*, and Andrew G. Clark*

Affiliation

Department of Molecular Biology and Genetics, Cornell University, Ithaca, USA
*Correspondence: sps257@cornell.edu, cf458@cornell.edu & ac347@cornell.edu

Abstract

The piRNA pathway is conserved to repress transposable element (TE) activity in the animal germline via a specialized class of small RNAs called piwi-interacting RNAs (piRNAs). piRNAs are produced from discrete genomic regions called piRNA clusters (piCs). While the molecular processes by which piCs function are relatively well understood in *Drosophila melanogaster*, much less is known about the origin and evolution of piCs in this or any other species. To investigate piC origin and evolution, we use a population genomics approach to compare piC activity and sequence composition across eight geographically distant strains of *D. melanogaster* with high quality long-read genome assemblies. We perform annotations of ovary piCs and genome-wide TE content in each strain. Our analysis uncovers extensive variation in piC activity across strains and signatures of rapid birth and death of piCs. Most TEs inferred to be recently active show an enrichment of insertions into old and large piCs, consistent with the previously proposed 'trap' model of piC evolution. By contrast, a small subset of active LTR families is enriched for the formation of new piCs, suggesting that these TEs have higher proclivity to form piCs. Thus, our findings uncover processes leading to the origin of piCs. We propose piC evolution begins with the emergence of piRNAs from a few specific LTR retrotransposon insertions that subsequently expand by accretion of other TE insertions during evolution to form larger 'trap' clusters. Our study shows that TEs themselves are the major force driving the rapid evolution of piCs.

Introduction

Animal genomes are parasitized by a horde of transposable elements (TEs) whose mutagenic activity can lead to sterility (Bingham et al. 1982; Bucheton et al. 1984). In the animal germline, the Piwi-interacting RNA (piRNA) pathway is a conserved small RNA-based mechanism regulating TE activity (Brennecke et al. 2007; Houwing et al. 2007; Lau et al. 2006; Grimson et al. 2008). piRNAs are 23-32 nucleotide RNAs produced from discrete loci called piRNA clusters (piCs) that guide effector Piwi proteins to silence TEs (Ozata et al. 2019). The piRNA pathway presents features of an adaptive defense system against TE invasion (Brennecke et al. 2008; Khurana et al. 2011; Yu et al. 2019; Aravin et al. 2007), but little is known about the processes driving its evolution. Many piRNA pathway genes encoding the proteins involved in TE silencing display signatures of adaptive evolution (positive selection) in several species' lineages (Simkin et al. 2013; Yi et al. 2014; Palmer et al. 2018), which may indicate adaptation to a rapidly changing TE sequence pool and new invasions (Cosby et al. 2019). However, little is known about the means by which piRNA-producing loci, piCs, originate and evolve in flies, or any other species.

piRNAs are produced from long non-coding RNA precursors that are transcribed from dispersed loci called piRNA clusters (Mohn et al. 2014; Brennecke et al. 2007). piCs make up 0.1-3% of the genome in fruit flies, mosquitoes, and mice, and are enriched for TEs and other repeats such as satellites, and sometimes host gene sequences as well (Brennecke et al. 2007; Houwing et al. 2007; Chirn et al. 2015; Ma et al. 2021; Chen et al. 2021). The best characterized function of piRNAs is to repress TEs. Since TE activity and composition vary significantly between and within species, TEs themselves may be important drivers of piC evolution, but this has not been thoroughly tested. TEs exhibit high diversity in their mechanisms of transposition and their genomic distribution (Sultana et al. 2017; Wells and Feschotte 2020) as well as spatial and temporal activity (Pasquesi et al. 2020; Lawlor et al.

2021; Chang et al. 2022). Hence, it is likely that piCs evolve through diverse mechanisms to repress newly introduced TEs.

The organization of piCs is best characterized in *Drosophila melanogaster*. The genome-wide piC landscape in the *D. melanogaster* ovary is comprised of tens of large (>10 kb) loci and hundreds of smaller (<10 kb) loci. It is also known that most large clusters (>10 kb) reside in pericentromeric and sub-telomeric regions. Larger pericentromeric clusters consist of tens to hundreds of diverse TE insertions, while the small clusters (<10 kb) often contain recent TE insertions (Shpiz et al. 2014; Baumgartner et al. 2022; Miller et al. 2023). The architecture and composition of some large clusters suggests a 'trap' model for the evolution of piCs, wherein TE insertions within clusters are selectively favored because of their production of piRNAs to repress all copies of the same active family (Bergman et al. 2006; Kelleher et al. 2018; Kofler 2019; Zanni et al. 2013; Brennecke et al. 2007). Over time, this process is predicted to result in piCs, which serve as the host's archive of past TE activity. This is consistent with observation of large piCs producing a bank of diverse piRNAs related to previously encountered TEs. It is important to note that recent studies reported bias in the ovarian piRNA sequence pool and piC loci composition toward younger TE insertions (Saint-Leandre et al. 2020; Gebert et al. 2021; Said et al. 2022). This indicates that evolutionary processes rapidly shape piC sequence composition to represent the most recently active TEs.

In contrast to the large canonical piCs, individual TE insertions can also function as piCs (Olovnikov et al. 2013; Le Thomas et al. 2014; Akkouche et al. 2017). Recent studies have demonstrated that piRNAs from large 'trap'-like piCs are neither essential for TE silencing, nor do they contribute the majority of piRNAs out of the total piRNA pool (Gebert et al. 2021; Genzor et al. 2021). Hence, piCs from individual TEs may produce a substantial abundance of piRNAs, which abrogates the need for active TEs to land into existing 'trap' clusters to come under the control of the piRNA pathway. While the 'trap' model of piC evolution has received

empirical support, its relative contribution to piC evolution in *D. melanogaster* is unclear, and the mechanisms underlying the origin and evolution of piCs remain broadly uncharacterized.

Results

Extensive variation in the genomic landscape of piCs

To quantify piC variation in *D. melanogaster* we generate a comprehensive annotation of active piCs in eight highly inbred strains. Seven of these strains are derived from natural populations of distinct worldwide origins and have publicly available long-read genome assemblies (Chakraborty et al. 2019). For each of these seven strains, we constructed and sequenced libraries of small RNAs isolated from ovaries of two biological replicates sampled 6 months apart (**Supplemental Fig. S1; Table S1).** In addition, we analyzed two ovarian small RNA libraries for the reference iso-1 strain, generated as part of two independent studies (Shipz et al, 2014 and Asif-Laidin et al, 2017). Each small RNA library is analyzed separately using the pipeline discussed briefly here and in more detail in Methods, which lists three methods to define piCs **(Supplemental Fig. S2).** The *restrictive* and *proTRAC* methods serve the purpose of discovering moderately to highly expressed piCs using uniquely and multi-mapping piRNAs respectively. The *permissive* method is carried out mainly to validate low to moderately expressed piCs detected by *proTRAC* using multi-mapping piRNAs. Both *restrictive* and *proTRAC* methods yield highly reproducible piC coordinates across replicates of each strain with >88% of total piC count shared between two replicates overlapping over >75% of their respective length **(Fig. 1A; Supplemental Fig. S3; Table S3)**. However, inter-strain pairwise comparisons reveal that pairs of strains share only an average of ~40% of their total piCs counts**.** The high-confidence set of piCs from *proTRAC* that either have high expression (>25) or were supported by uniquely mapping piRNAs, along with all piCs from the *restrictive* method for each replicate, are combined to create a replicate-specific ‘master list’ of piCs.

Genome-wide visualization of piC annotations in iso-1 coordinates across chromosomes reveals variability in the piC landscape across the 8 strains **(Fig. 1B; Supplemental Table S4)**. In aggregate, the total amount of genomic DNA covered by active piCs in each strain ranged from 4.8 Mb to 6.3 Mb **(Fig. 1C)**, encompassing 3.4 to 4.8% of their respective genome assemblies (**Supplemental Fig. S4C**). While their piC landscape is broadly similar in terms of being denser within peri-centromeric and telomeric heterochromatic regions (characterized by low mappability scores) compared to euchromatic regions, it is readily apparent that many individual clusters are present in only one or a few strains, even within these heterochromatic regions. Often smaller but not exclusively euchromatic piCs are even more variable across strains despite being characterized by higher mappability scores **(Fig. 1B)**. Thus, from this broad-scale view, it appears that the total span of the genome occupied by piCs within each strain is largely similar, but the positions of piCs are highly variable across strains.

Figure 1. Inter-strain variability of piCs in *D. melanogaster* strains. (*A)* Cross strain and replicate overlap (% of total piCs count) of independently predicted piCs for each small RNA library using the restrictive method. Inter-strain comparisons are done after piC were remapped to iso-1 coordinates and overlap >1 bp between two piCs is deemed as shared. Intra-strain comparisons are done similarly but in native genome assembly. *(B)* Genome-wide distribution of lifted over piCs in 7 DSPR strains and reference iso-1 strain. Bars along the circumference represent presence of piCs in 10 kb bins for each Chromosome. The outermost bar plot is piRNA mappability scores, followed by iso-1 piCs, followed by piCs of 7 DSPR strains. *(C)* Combined genomic piC size predicted from each replicate small RNA library independently in respective genome assemblies. *(D)* Combined piC length predicted independently for each strain per small RNA library. *(F)* Population frequency of piCs quantified after liftOver to the reference iso-1 genome. *(G)* piC length distribution by population frequency in kilobase-pairs (kb) quantified after liftOver to the iso-1 genome.

Abundant strain-specific and strain-biased piCs

To quantify the frequency of piCs across the 8 strains, we scored the overlap of piCs predicted independently for each of the 8 strains using the master list coordinates. To account for changes in size of piCs among strains, we required a minimum positional overlap of only 1 bp for piCs to be considered shared between strains. Even when using this non-conservative criterion, we found that 568 (*restrictive*) to 906 (*proTRAC*) of piCs are active only in a single or a few strains confirming that each strain has a unique piC landscape **(Figure 1D).** The results are similar whether we used the piC predictions of the *restrictive* and *proTRAC* methods separately or the combined master list **(Fig. 1D; Supplemental Table S2).** All strains contained 35-60 piCs that are strictly unique to that strain and another ~30 piCs (<10%) that could not be lifted-over and therefore are likely to be strain-specific also **(Supplemental Fig. S4A)**. Thus, we can conservatively estimate that each strain possesses >50 piCs that are not shared by any of the other 7 strains examined. In addition, 142 and 69 piCs (*restrictive* pipeline) are shared between two and three strains respectively. All such piCs, shared by 4 strains or less are together termed as 'rare' piCs. Rare piCs are not only extremely abundant but also exhibit significant piRNA expression ranging from 10-25 RPM, which is comparable to previously described canonical piCs like *80EF* and the *traffic jam* 3'UTR **(Fig. 2; Supplemental Figure S5)**. Additionally, despite their small size, in aggregate, strain-specific piCs contribute a substantial portion of the total genomic span of piCs (average of ~1 Mb) and 15-20% of the total piC genomic length of each strain **(Supplemental Fig. 4B).**

Next, we examined the relationship between the size of piC and their level of sharing across strains. First, we note that piC length predicted from each library is similar, with median from 5.7 kb to 7.5 kb **(Supplemental Fig. S4C)**. We find that piC size is positively correlated with the level of sharing across strains, and this correlation holds true for all prediction methods (Pearson r = 0.56 for proTRAC, 0.58 for restrictive, and 0.76 for master-list, *p-value* < 2.2×10^{-16})

(**Fig. 1E)**. We also verified this correlation with differing piC overlap parameters to compute sharing across strains **(Supplemental Fig S11)**. In other words, piCs detected in a single or a minority of the strains (rare piCs) tend to be smaller (2-10 kb) than those shared by most of the strains (common piCs). If we posit that rare piCs represent evolutionarily younger piCs than common piCs, this relationship suggests that piCs are born relatively small and increase in size as they get older. Alternatively, larger piCs may be more evolutionarily stable than smaller ones. We note, however, that even large piCs can still be variable in activity across strains. For example, large well-known piCs like *42AB* and *38C* are still only active in 6 or 7 of the 8 strains (see below). Taken together, these results suggest that ovarian piCs are extremely labile and poorly conserved in activity across *D. melanogaster* strains.

Extensive variability in piRNA expression of piCs

To illustrate the differences in activity of piCs among the 8 strains, we examined the piRNA coverage profiles for *42AB* and *38C*, two large dual-stranded piCs and two small piCs with varying activity across strains in *76CD* cytogenetic region **(Fig. 2)**. The *42AB* piC has been extensively documented for its high uniquely-mapping piRNA expression (Brennecke et al. 2007; Klattenhoff et al. 2009). We present normalized coverage of uniquely mapping piRNAs to per million miRNAs **(Fig. 2A)** for the respective *42AB* assemblies from both small RNA library replicates of four strains. Additionally, to examine differences in read coverage due to mappability, theoretical mappability scores are visualized along the length of the cluster in 100 bp bins. Strains A1 and A7 have severely reduced (>20-fold) piRNA expression levels throughout *42AB* compared to the other strains, whereas iso-1 and B6 exhibit *42AB* piRNA abundance reported in other studies in the literature (Brennecke et al. 2007; Klattenhoff et al. 2009). Congruence in piRNA abundance between miRNA and sequencing depth normalization strategy is also confirmed, shown for *38C* **(Supplemental Fig. S5)** and additional small piCs for strain B6 **(Supplemental Fig. S6)**. Similarly, *38C* – a highly active dual-stranded piC in iso-1,

exhibits significant variability in uniquely-mapping piRNAs across strains **(Supplemental Fig. S5)**. Since *42AB* and *38C* are active in 6 out of 8 strains, it is most parsimonious to conclude that these piCs are relatively old but have lost activity in a subset of strains.

Two piCs in *76C and 76D* cytogenetic region are shown because each exhibits different levels of sharing across strains. *76C* is a ~15 kb piC, likely uni-stranded, exclusively active in strains A7 and B3, whereas *76D* is a dual stranded piC ranging from 6 kb to 20 kb in size but active in all the 6 strains shown **(Fig. 2B).** *76E*, a uni-stranded piC with low piRNA abundance is considered active in all strains except A2. Comparison of such syntenic piCs between strains in their native genome assemblies provides validation of the variable activity of piCs across strains presented earlier from annotation pipelines **(Fig. 1B).** In addition, we also note that while genomic assemblies of *42AB* from different strains are similar in size, there are structural rearrangements, which may contribute to differences in sequence composition of same piC across strains **(Fig. 2C; Supplemental Fig. S8, S9)**.

Figure 2. Natural variation in expression of uniquely mapping piRNAs from *42AB* and *76C and D*. *(A)* Uniquely mapping piRNA abundance profiles of *42AB* piCs for the 4 strains with two small RNA library replicates. Y-axis values are piRNA reads for 100 bp bins per million miRNA reads. Mappability scores (0-1) is shown for 100 bp bins of each respective *42AB* genomic assembly in heatmap. *(B)* Uniquely mapping piRNA expression profile of *76C, D and E* piCs for the 6 strains with two small RNA library replicates. Y-axis values are piRNA reads for 100 bp bins per million miRNA reads. Mappability scores (0-1) are shown for 100 bp bins of each respective *76C, D, and E* genomic assembly in heatmap. *(C)* Ribbon plot of multiple-sequence alignment for *42AB* assembly in 4 strains. Only alignments of >80% identity and >1kb in length are shown.

Structural variation in piCs supports aspect of the 'trap' model

To understand the mutational processes underlying the changes in structure and sequences of piCs among strains, we examine the contribution of inter-strain structural variants (SVs), namely insertions and deletions (indels). We detect indels for each of the strains relative to the iso-1 reference strain (Chakraborty et al. 2019; Solares et al. 2018). We map raw long sequencing reads for each strain to the iso-1 genome and call indels using three independent SV callers (see Methods). Indels from all strains were then collapse to construct a list of unique SVs that consisted of putative 2274 insertions and 4409 deletions. Indels are then polarized into 'true' insertions and deletions by comparison of each variant to *D. simulans* and *D. sechellia* reference strains, which enable inference of the ancestral state (see methods, **Supplemental Fig. S10 A,B**). Polarization led to loss of ~55% of indels as the ancestral or derived state of the loci could not be determined due to conflicts in calls between the two outgroup species. After this filtering, 1183 insertions and 1873 deletions are retained for analysis **(Supplemental Table S6).**

We examine the size distribution of indels overlapping piC and non-piC regions of the genome to then test the predictions for indels associated with piC variation. First, genome-wide, insertions range from 30 bp to 91 kb with a median of 612 bp, while deletions range from 30 bp to 7.6 kb with a lower median of 208 bp compared to insertions **(Fig. 3A)**, which is consistent with previous SV profiling of *D. melanogaster* strains (Dopman and Hartl 2007; Zichner et al. 2013; Huang et al. 2014). However, insertions overlapping piCs have a median length of 2.2 kb, whereas insertions nonoverlapping piCs have a smaller median length of 512 bp (Kruskal-Wallis test, χ^2=10.812, df=1, *p*-value = 0.001). Meanwhile, deletions overlapping piCs have a similar length distribution than those nonoverlapping piCs (Kruskal-Wallis test, X^2=0.72404, df= 1, *p*-value = 0.39) **(Fig. 3B)**. We also compare the length distribution of indels in piCs grouped by the number of strains with which they are shared. We find that strain-specific or rare piCs (shared by less than 4 strains) are associated with relatively large insertions (median length of 5.2 kb),

whereas common piCs (shared by more than half of the strains) have a median insertion length of less than 1 kb **(Fig. 3C)**. In sum, rare piCs are uniquely associated with relatively large insertions, which is consistent with the idea that these piCs emerged from recent TE insertions.

Next, we test whether piCs are enriched for indels relative to the rest of the genome. To do this, we compared the indel counts overlapping piCs for each of the piC frequency categories with those expected based on 1000 sets of randomly shuffled indels. We find that deletions are significantly enriched in common piCs, but not in rare piCs **(Fig. 3D)**. Insertions are strongly enriched both in rare and common piCs **(Fig. 3E)**. These results may be confounded by the location of many piCs within constitutive heterochromatin, where the rate of SVs is generally high (Chakraborty et al. 2021; Montgomery et al. 1991). However, we find that only ~28% of all piCs lie within constitutive heterochromatin boundaries of the reference genome assembly and indels are significantly enriched in piCs even when we compare them to heterochromatic regions **(Supplemental Fig. S7)** (Riddle et al. 2011).Thus, the SV enrichment we observe within common piCs is unlikely to be solely driven by their location within constitutive heterochromatin. Overall, we conclude that piCs are subject to a high rate of structural genomic change relative to the rest of the genome, which likely contributes to their rapid evolutionary turnover. Additionally, we find that common piCs are enriched for both insertions and deletions, which is consistent with these clusters evolving as ‘traps’. By contrast, rare piCs are only enriched for insertions, which supports the notion that these are generally young clusters born from recent, singleton TE insertions.

Figure 3. Common piCs exhibit ‘trap’ like sequence turnover. *(A)* Observed counts of indels genome-wide and overlapping with piCs. *(B)* Length distribution of indels genome-wide and overlapping with piCs. Significant differences are shown from Kruskal-Wallis test comparisons. *(C)* Length distribution of indels overlapping piCs grouped by the number of strains they are shared by (i.e., population frequency). *(D&E)* Enrichment analyses of deletion (DEL) and

insertions (INS) variants in piCs carried out. Variants overlapping with piCs of differing population frequencies is compared to the expected mean overlap counts genome wide. Variant calls were shuffled 1000 times (p -value <0.05=*, <0.005=**).

TE re-annotation of each strain uncovers ~3 Mb of unannotated TE DNA.

While our analysis of SVs within piCs supports that these events are important drivers of piC evolution, it does not directly address the role of TE activity. To assess the contribution of TEs to the composition and changes in the activity of piCs across strains, we carried out *de novo* annotation of TEs in each of the strain genome assemblies. This was necessary because many TE consensus sequences present in the reference TE library for *D. melanogaster* (FlyBase release 2019_05) were discovered and curated more two decades ago using primarily the iso-1 and Oregon-R strains (Bartolomé et al. 2002a; Kaminker et al. 2002; Bowen and McDonald 2001). However, recent advances in long-read sequencing technology have provided a means to obtain a more unbiased view of the repetitive landscape of *Drosophila* genomes, revealing novel TE families (Ellison and Cao 2020; Rech et al. 2022; Han et al. 2022). We developed a TE annotation pipeline based on RepeatModeler2 (for discovery) (Flynn et al. 2020; Smit 1999), RepeatMasker (for annotation) and additional tools to distinguish novel TEs from known TEs and curate a comprehensive TE library for the 8 strains used in this study **(Fig. 4A,** See methods**)**.

TE family sequence assemblies by RepeatModeler2 for each strain were aligned to the reference TEs using the 80-80-80 rule (Wicker et al. 2007). RepeatModeler2 sequences already present in the reference TE library were removed. Next, the remaining RepeatModeler2 sequences were used to re-mask the genomes to examine them for novel TE families. Over 5000 insertions (>500 bp in size, median of 676 bp), very similar (<5% divergence) to their respective RepeatModeler2 family consensus are discovered for each strain (example of strain B6 in **Fig. 4B**). These novel insertions resulted in masking of an additional 2.5 Mb to 4 Mb in

each genome assembly that would have been missed or mis-annotated as highly diverged insertions by masking with only the reference TE library **(Fig. 4C)**. In summary, a refined and comprehensive TE library was created with a combination of 129 reference TE consensus sequences and 47 uncharacterized consensus **(Supplemental Table S7)** sequences that capture all TE insertions genome-wide and reflect their relative age.

Figure 4. *de novo* TE annotation uncovers ~3 Mb of hidden TEs and reveals strong associations of young LTR TEs with piCs than any other TE subclasses. *(A)* TE annotation pipeline using RepeatModeler2 and RepeatMasker to create the comprehensive TE library. *(B)* Abundance of extremely similar and long TE insertions from RepeatMasker output of strain B6 using novel TE consensus library. *(C)* Differences in million base-pairs (Mbps) masked in RepeatMasker results using novel-only, reference-only, and combined TE library. *(D)* Divergence estimates for all defragmented iso-1 insertions (>200 bp) from RepeatMasker output. Insertions with >1 bp overlap with master-list iso-1 piCs are considered piC overlapping. Difference between groups is tested by Wilcox ranked-sum test. *(E)* Terminal branch length for all iso-1 insertions from *Ty1*/*Copia* and *Tc1*/*Mariner* superfamilies from maximum likelihood trees. *(F-H)* Maximum likelihood trees constructed from all defragmented insertions for *blood*, *Tc1-2*, and *G-element* families and the inset shows terminal branch length quantification. Difference between groups is tested by Wilcox ranked-sum test; p-value <0.05=*, <0.005= **, <0.0005=***.

piC TE composition is represented by younger LTR insertions than any other TE subclass.

To understand the co-evolution of TE and piCs, we correlated TE insertion age and piC sharing across strains (as proxy for piC age). Using the new TE library described above, we sought to compare the age and composition of TEs within piCs to that of the rest of the genome. To infer

the age of each family, we use the median sequence divergence of each insertion to their family consensus. To examine TE composition, we grouped TEs into the major subclasses and superfamilies represented in *Drosophila*: non-LTR retrotransposons (LINE), LTR retrotransposons (*Ty1*/*copia*; *Ty3*/*mdg4*; *BEL*/*Pao* superfamilies), Rolling Circle (RC) transposons, and cut-and-paste DNA transposons. We found that TE copies from all three LTR superfamilies are significantly younger in piCs than non-piC regions **(Fig. 4D)**. Conversely, TE copies from LINE, RC and DNA subclasses are not significantly different in age in piCs than in non-piC regions. To test these results using an independent method to date insertions, we built phylogenetic trees from all copies for one LTR superfamily (*Ty1*/*copia*) and one DNA transposon superfamily (*Tc1*/*mariner*) and used terminal branch lengths to estimate their relative age (Carr et al. 2012). We chose these superfamilies because they are of moderate abundance and therefore manageable for multiple sequence alignments and phylogenetic analyses. The results of these analyses yielded the same trend observed genome-wide using sequence divergence from consensus sequences whereby the *Ty1*/*copia* LTR retrotransposons (*n*=135) overlapping piCs are significantly younger than nonoverlapping ones, while *Tc1*/*mariner* (*n*=89) DNA transposons show no such bias **(Fig. 4E)**.

To examine whether these trends hold at the level of individual TE families, we selected one family with moderate copy number from the LTR, LINE and DNA subclass and compared the age of piC-overlapping and nonoverlapping copies within each family. As a representative *Ty3*/*mdg4* LTR superfamily, we analyzed *blood*, a family with 63 copies in the iso-1 strain that is known to be transpositionally active (Bingham and Chapman 1986; Kofler et al. 2015). Consistent with the trend observed at the level of the LTR superfamily, we found that 43 out of 63 *blood* insertions are associated with piCs. Most of these are very recent insertions with median terminal branch length of <0.002, which is significantly shorter than that of insertions not overlapping with piCs (Wilcoxon rank sum test, *p*-value = 0.014) **(Fig. 4F)**. In other words, piC

overlapping *blood* insertions are significantly younger than the nonoverlapping ones. As a representative of the *Tc1/mariner* superfamily of DNA transposons, we analyze *Tc1*-2, a family with 35 copies in the iso-1 genome. Consistent with the trend observed at the level of the entire superfamily, the age of *Tc1-2* copies overlapping piC is not significantly different than that of non-piC overlapping copies (Wilcoxon rank sum test, *p*-value = 0.903) **(Fig. 4G)**. Analyzing the *G*-element LINE family, which counts 35 copies in iso-1 and is still active (Di Nocera et al. 1986), we found that the age of piC-overlapping copies is not significantly different from nonoverlapping copies (Wilcoxon rank sum test, *p*-value 0.855) and the youngest *G*-element insertions according to terminal branch length do not overlap piCs **(Fig. 4H)**. Taken together, these results suggest that young LTR retrotransposon insertions tend to be enriched in piCs, but this trend is not observed for other TE subclasses and superfamilies.

A small subset of active LTR retrotransposon families give rise to young piCs.

To test the central prediction of the 'trap', where recent transposition events from active TEs must be enriched in piCs, we established a set of non-reference TE insertions in each DSPR strain using the raw long read data available for each. Briefly, we applied TLDR (a long-read TE insertion detection tool) (Ewing et al. 2020) with a cut-off of at least 2 supporting reads per 10× genome coverage to remove false positives and enrich for germline insertions (see Methods). Using these parameters, we identified 285 to 857 non-reference TE insertions in each of the 7 DSPR lines but only 75 insertions for iso-1, which is expected since the reference genome is also derived from the iso-1 strain. Presumably, the 75 non-reference insertions for iso-1 reflect the use of different isolates for the reference genome assembly and for the long-read data. Further clustering and parsing of all non-reference insertions across the 8 strains resulted in a list of 3545 unique TE insertions of at least 200 bp in length **(Supplemental table S8)**. These insertions belong to 165 of the 184 different families in our TE library. Ninety-four of these 165 TE families were classified as "active" because each of these included at least 10 non-reference

insertions shared by no more than 2 strains, while the other 98 families were classified as "inactive" **(Supplemental Table S9).**

We use this compendium of insertions to test whether active TEs are significantly enriched within piCs using a binomial test to compare the observed overlaps with piCs to the average overlaps expected from 1000 random reshufflings of TE insertions (see Supplemental Methods) (Kapusta et al. 2013). This analysis revealed that 7 active TE families are significantly enriched in piCs, while 1 active and 10 inactive families are significantly depleted in piCs **(Fig. 5A).** All the inactive TE families that are significantly depleted in piCs belong to either DNA or LINE subclasses, with the exception of one active family depleted in piCs was *17.6*, a *Ty1/copia* superfamily member (Inouye et al. 1986) with 79 non-reference insertions. These results are consistent with the prediction of the 'trap' model that piCs are enriched for active TE families but are also composed of inactive families.

Next, we sought to distinguish family-level enrichment of TE insertions within rare, recently arisen piCs (smaller piCs (<10 kb) shared by no more than 3 strains) and within common larger 'trap'-like piCs (>10 kb, shared by at least 5 strains). To increase statistical power for this analysis, we use all TE sequences annotated by RepeatMasker in each genome, instead of only non-reference insertions. For iso-1, we find that only 4 TE families are significantly enriched within young piCs. All 4 are LTR retrotransposon families of the *Ty3/Mdg4* superfamily (*blood, Dm-412, flea, Stalker-2*) and all except *flea* belong to the *Mdg1* lineage **(Fig. 5C,D)** (Bertocchi et al. 2020; Costas et al. 2001). All four families are also classified as active in this study as well as previous studies that examined TE insertion frequency among *D. melanogaster* populations (Kofler et al. 2015; Kelleher and Barbash 2013). By contrast, we find that numerous active and inactive families from all TE subclasses are significantly enriched in large and common "trap-like" piCs, largely representative of the overall TE landscape of *D. melanogaster* **(Fig. 5C,D)**. In the A1 genome, 20 TE families are significantly enriched in rare piCs. Again, these are

predominantly LTR retrotransposons (14 families), but 4 DNA transposon families and 2 LINE families are also significantly enriched. **(Fig. 5E,F)**. *1360*, a DNA transposon, is one of the 4 enriched DNA transposon families in A1 rare piCs but not in iso-1. It is possible these differences arise from changes in TE activity, number of recent insertions, or strain background impact TE's readily conversion into active piCs. *blood* insertions are neither enriched nor depleted in common piCs of both strains. Taken together, these analyses yield a contrasting portrait of TE composition in the two major types of piCs.

Figure 5. Insertions of only few active LTR families associates with rare piCs. *(A)* Scatter plot of non-reference TE insertion counts and mean population frequencies of 176 TE families. Black dots highlight selected TEs classified as 'active', whereas gray dots are deemed 'inactive'. *(B)* Enrichment analyses of non-reference insertions of TE families in master-list piCs using random shuffling of non-reference TE insertions. Y-axis *P*-values are from binomial tests conducted to compare observed counts to expected average overlaps of non-reference TE insertions to piCs for each family. *(C-D)* Enrichment analyses of TE families in master-list rare and common piCs of A1/Can-S strain using random shuffling. All TE insertions (>200 bp) of strain iso-1 genome are randomly shuffled with constant piCs coordinates grouped into rare (shared by 4 strains or less) and common piCs (shared by 5 strains or more). *P*-values on y-axes are from binomial tests conducted to compare observed counts to expected average overlaps of TE insertions to piCs for each family. Statistically significant families are labelled.

TE composition of piCs captures distinct steps in piC evolution.

To further illuminate the evolution of piCs, we analyzed how the overall age and distribution of TEs of piCs change as they become more frequent in the population, and presumably older. We plotted mean percent divergence of individual TE insertions to their consensus sequences (a measure of TE age) across piCs and their flanking non-piC regions for each piC frequency class (strain-specific or shared by 2-8 strains). We find that the divergence of TE insertions in rare

piCs (shared by 3 or fewer strains) is markedly lower (3.5-5%) than in their flanking regions (10-15%) **(Fig. 6A)**. In addition, the divergence of TE copies within piCs increases gradually with the frequency of the piCs to the extent that for the most common piCs (shared by 7 and 8 strains) the mean percent TE divergence is only slightly lower than in their flanking genomic regions. This apparent increase in average age of TE insertions as piCs become more frequent provides weight to the inference that more common piCs represent evolutionarily older clusters relative to those that are strain-specific or rare. It also suggests that piCs are born from singleton TE insertions and grow by gradual accretion of TEs over time.

To further test this idea, we examine how TE coverage within piCs and surrounding regions change as piCs become more common and presumably older **(Fig. 6B)**. First, we observe that piCs generally exhibit significantly higher TE coverage than their flanking genomic regions. Second, we find that strain-specific piCs, which presumably represent the youngest piCs, exhibit high mean TE coverage in the middle at >60% (on average 60 out every 100 bp is comprised of TE sequence), which drops at the edges of piC coordinates to <20% **(Fig. 6B)**. In contrast, more common piC groups exhibit consistently higher TE coverage across their entire length. This pattern is consistent with a birth and growth process where a piC emerges from individual TE insertion, but piRNA production spreads to flanking TE insertions as they insert near or within the piC.

Figure 6. Age and distribution of TEs provide a portrait of intra-specific piC evolution. (A) TE age (by % divergence) is plotted for piCs of increasing population frequency from left to right. Summary profile plot of mean percent TE insertion divergence in 500 bp windows in scaled piC regions and flanking non-piC regions of the reference genome. Each group represents piCs that are shared by 1/8 to 8/8 strains. (B) TE content of piCs of different stages are plotted. Summary profile plot and heatmap of mean TE coverage in 500 bp windows in scaled piC regions and flanking -/+ 15 kb non-piC regions of the reference genome. Scaled piC

regions are marked by start and end labels at the bottom of the heatmap. (C) New unified model of piC evolution - “birth-and-death” is shown in four steps from left to right.

Discussion

To study piC evolution at fine-scale resolution in *D. melanogaster*, we used population genomics methods to characterize piC variation across eight inbred strains. A crucial asset was the availability of high-quality genome assemblies for these strains (Chakraborty et al. 2019). This enabled us to produce *de novo* annotation of piCs for each strain from mapping inferred piRNAs from ovarian small RNA libraries we constructed and sequenced for two biological replicates sampled six months apart. Our piC annotations for the two replicates exhibited high reproducibility with >88% of piCs annotated in one replicate found in the second replicate **(Fig. 1A; Supplemental Fig. S3)**. Also, to understand variation in sequence composition and age of piCs, it was necessary to produce libraries of TE consensus sequences representative of the eight strains analyzed here. By performing *de novo* discovery and re-annotation of TE families for each genome, we identified 47 novel TE families **(Fig. 4)**. While further investigation is required to examine their evolutionary origins and relationship to known TE families, it appears that many of the novel TE families we annotated were highly diverged from known families and often “hidden” in highly repetitive regions that would likely be poorly assembled in short-read genome assemblies. These results stress the benefits of high-quality genome assemblies and the necessity to perform *de novo* TE discovery when new strains or geographical isolates are considered. This is true even for model species like *D. melanogaster,* where TEs have been extensively cataloged, because previous TE identifications were mostly based on a single reference genome. Robust annotation of piCs and TEs allowed us to compare the activity and TE composition of piCs across strains and derive general principles piC origin and evolution.

Our findings recapitulate aspects of the ‘trap’ model of piC evolution. First, we find enrichment of insertions of active TE families and depletion of inactive families in master-list piC regions **(Fig.**

5B). In addition, we also find that 'trap'-like large piCs are enriched for diverse TE families, which are relatively younger in piCs than flanking non-piC regions **(Fig. 5D,6A)**. This is consistent with the idea that 'trap' piCs represent an archive of past TE activity which biases piRNA production toward recently active TE families. Consistent with previously reported variation in piC activity (Rozhkov et al. 2010; Zhang et al. 2020; Ellison and Cao 2020), we find extensive intraspecific variation in piC activity, even for large 'trap'-like piCs. For example, several *42AB* and *38C* clusters show significant loss in piRNA production in one or multiple strains **(Fig. 2; Supplemental Fig. S6)**. What could cause the loss in piRNA production from large heterochromatic piCs? Our current lack of understanding of the *cis*-regulatory requirements for piC activity makes it difficult to determine whether changes in piRNA production are caused by genetic or epigenetic changes in the piCs. Consistent with previous studies (Wierzbicki et al. 2021b; Ellison and Cao 2020), we observe considerable structural variation in large peri-centromeric piCs **(Supplemental Fig. S7)**. It is possible that such structural changes result in changes in piRNA production, but further studies are needed to elucidate the mechanisms by which large and seemingly stable piCs lose their activity.

What can TE composition of piCs tell us about the coevolution of TEs and piCs? As previously reported (Wierzbicki et al. 2023; Kofler 2020), we find that diverse TE families (from all subclasses) are enriched in large, common piCs **(Fig. 5)**. This enrichment may be explained by selection against new TE insertions in gene-rich euchromatic regions, which leads to accumulation of TEs in heterochromatic regions, where purifying selection is also weak (Blumenstiel et al. 2002; Schrider et al. 2013; Charlesworth and Langley 1989; Dolgin and Charlesworth 2006; Bartolomé et al. 2002b). However, our genome-wide analysis revealed that SV enrichment in common piCs cannot be completely explained by their overlap with heterochromatin **(Supplemental Fig. S8)**. Thus, the enrichment of diverse elements within large

common piCs may be driven in part by high tolerance of deleterious effects of TEs and/or by selection for their repression.

Are particular TEs prone to give rise to piCs? To answer this question, we tested for enrichment of individual TE families within rare piCs, which likely represent young clusters. Only a small set of retrotransposon families are enriched within such clusters and most belong to the *mdg1* subclade of LTR retrotransposons such as *blood* and *Dm-412*. **(Fig. 5)**. Why would these elements be prone to seed new piCs? There are two possibilities, which are not mutually exclusive. One possibility is that these TE families are readily licensed as piCs from maternally inherited piRNAs, thereby giving rise to emergence of dispersed piC loci following recent transposition events (Le Thomas et al. 2014; Shpiz et al. 2014). Alternatively, it may be linked to their propensity to produce double-stranded RNA (dsRNA) and endogenous siRNAs, as endo-siRNAs production has been associated with the formation of piCs from transgenes (Olovnikov et al. 2013; Akulenko et al. 2018). Many LTR (including *blood*) and non-LTR retrotransposons are known to possess bi-directional promoters that can result in the formation of dsRNAs that stimulate the production of siRNAs (Watanabe et al. 2008; Russo et al. 2016; Hung and Slotkin 2021). It is possible that the inherent capacity of some retrotransposons to produce dsRNAs nucleate the formation of piCs. This idea has received support in a recent study showing that endogenous siRNA production precedes piRNA cluster formation, and maternal inheritance of these siRNAs is required for licensing of piRNA clusters (Luo et al. 2022). However, we cannot rule out that these rare TE-specific piCs arose exclusively from maternally inherited piRNAs as piRNAs from such piCs could not be differentiated from the piRNAs produced from insertions in large common piCs **(Supplemental Fig S12)**.

In summary, a picture is emerging in which the rapid evolution of piCs among *D. melanogaster* strains is driven by the activity of TE families themselves. We propose a ’birth-and-death’ model of piC evolution, which combines components of the conventional ‘trap’ model and features of

piCs found in this study (Bergman et al. 2006; Khurana et al. 2011; Shpiz et al. 2014; Gebert et al. 2021). In this model **(Fig 6C)**, we posit that piCs form frequently throughout the genome, often from recent TE insertions, likely through maternally deposited piRNAs, with certain LTR retrotransposon families making a stronger contribution to seeding new piCs. Newly emerged piCs may increase in frequency and size due to natural selection or drift, depending on factors such as their propensity to trigger genomic autoimmunity (Blumenstiel et al. 2016; Huang et al. 2022), ectopic recombination (Petrov et al. 2003; Sentmanat and Elgin 2012),and the establishment of a chromatin environment conducive for piRNA production (Le Thomas et al. 2014). Over time, these stabilized clusters may grow by 'trapping' additional TE insertions, which will eventually result in large heterochromatic clusters such as *42AB*. Due to the host's limited capacity to maintain such piCs without incurring a fitness cost, those clusters that lack piRNAs targeting active TEs may gradually lose activity or become dispensable (Gebert et al. 2021). We can only speculate about the mechanisms that lead to the loss of activity of large piCs, but the rapid turnover of piC loci provide some possibilities. Due to re-establishment of female piC loci every generation from maternally deposited piRNAs, it is possible that a large piC may not get licensed if it only contains old insertions from inactive TE families which are poorly represented in maternally deposited piRNAs. Since the maternally deposited piRNA pool is expected to diverge from generation to generation to incorporate new TEs, it may not contain enough homologous piRNAs for licensing of older piCs that lack recently active TEs. This seems consistent with high piRNA abundance of *42AB* in 5 of the 8 strains, low in 2 of the 8 strains and extremely low in one strain. The most parsimonious conclusion from such continuous variation can be drawn that *42AB* activity is on the decline in *D. melanogaster* populations. Further studies are warranted to test this "birth-and-death" model. Our study provides a first in-depth view of piC evolution in *D. melanogaster* that is likely to stimulate other comparative studies of piRNA evolution.

Materials and Methods

Fly stocks

DSPR founder stocks of A1 (b1_paired), A2 (b3841_paired), A4 (b1_3852), A7 (t7_paired), B3 (b3864_paired), B6 (t1_paired) and Oregon-R were a gift from Anthony Long (UCI). All stocks were maintained on standard cornmeal medium at 22°C under a 12-hr day/night cycle.

Small RNA library construction and sequencing

Small RNA libraries were constructed by size fractionation on urea-polyacrylamide gel electrophoresis as described in Ma *et al.* (2021) and additional details are provided in Supplemental Methods. All libraries were quantified using Qubit 3.0, pooled into replicate-1 and replicate-2 groups, and analyzed on Agilent Bioanalyzer. Single-end 75 bp Illumina sequencing was carried out for all libraries on NextSeq 500 at Cornell Biotechnology Resource Center.

piC annotation

Active piRNA cluster (piC) annotation was conducted independently for each replicate of each strain using a custom pipeline adapted from previously described methods (Mohn et al. 2014; Rosenkranz and Zischler 2012). Detailed annotation steps of each pipeline are provided in Supplemental Methods and outlined in **Supplemental Fig. S2**.

Structural variation detection and filtering

Raw long reads for the 7 DSPR strains, the iso-1 reference strain, *D. simulans* (wxd1) and *D. sechellia* (sech25) were mapped to the *D. melanogaster* iso-1 release 6 (GCA_000001215.4) without the Y Chromosome with minimap2.1 map-pb --N3 and resulting sam file was converted to BAM and sorted (Li 2018; Li et al. 2009). Three structural variant (SV) callers – Sniffles-2.0 (Smolka et al. 2024), cuteSV-1.0.13 (Jiang et al. 2020), and svim-2.0 (Heller and Vingron 2019) were used for SV detection. Filtering, genotyping, and collapsing of SV calls are detailed in Supplemental Methods. Filtered and polarized SV calls are reported in **Supplemental Table S5**.

de novo TE annotation

To create a comprehensive and accurate representative TE library representing the TE insertions contained in the 8 strains, *de novo* TE annotation was conducted using several computational tools. Summary of all major steps is presented in a flow-chart in **Figure 4A**. Briefly, canonical FlyBase TE consensus sequences were filtered to include only TEs that best represent the TE insertion landscape of each strain using RepeatMasker-4.1.0 results(Smit 1999; Larkin et al. 2021). FlyBase TE families with at least 3 copies of >200 bp and <1% divergence was retained. This resulted in a reference library for the 8 strains with 129 TE families. Next, RepeatModeler2 was run on the 7 DSPR genomes and reference iso-1 strain followed by removal of non-TE repeats like tRNA, satellites, rRNA etc., as well as TE sub-families using bash scripts (Flynn et al. 2020). Putative novel TE family consensus identified (See Supplemental Methods) and each genome were re-masked with combined TE library (novel and reference TEs).

Non-reference TE insertion detection from long reads

Raw long reads utilized in SV detection were also used for unique TE insertion analyses. Reads (>1 kb) were mapped to iso-1 reference genome (without Y-linked contigs and all contigs<20 kb) using minimap2.1 default parameters and resulting sam file converted to BAM file, sorted and indexed using SAMtools (Li 2018; Li et al. 2009). TLDR, a *de novo* TE detection program (Ewing et al. 2020)was run for each strain using comprehensive *D. melanogaster* TE library curated in this study. High confidence TLDR insertion calls for all 8 strains are reported in **Supplemental Table S7**.

Data access

All raw and processed sequencing data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA1013381. Relevant source data and methods for figures are available as Supplemental

549 Material. Additional raw data files, TE consensus sequences and scripts are available in
550 Supplemental Code and at https://github.com/kerogens101/Dmel_piC.

551 **Competing interest statement**

552 The authors declare no competing interests.

553 ## Acknowledgements

554 This work was supported by National Institute of Health grants - R01-GM119125 to A.G.C and
555 R35-GM122550 to C.F. S.P.S is supported by the Distinguished Scholar Award from the Cornell
556 Center for Vertebrate Genomics. We thank Jullien Flynn for her assistance in RepeatModeler2
557 runs and consensus filtering and Michelle Stitzer for recommendations on structural variant
558 analysis. We also thank Dan Barbash, Justin Blumenstiel and Andrew Grimson for helpful
559 discussions on the analyses and members of the Clark and Feschotte laboratories for valuable
560 feedback on the manuscript.

561 *Author Contributions*: S.P.S performed all experiments. S.P.S performed all computational
562 analysis and created all figures with input from A.G.C and C.F. S.P.S, A.G.C and C.F wrote the
563 manuscript. Initial project conception was by S.P.S and A.G.C.

564 ## References

565 Akkouche A, Mugat B, Barckmann B, Varela-Chavez C, Li B, Raffel R, Pélisson A, Chambeyron
566 S. 2017. Piwi Is Required during Drosophila Embryogenesis to License Dual-Strand piRNA
567 Clusters for Transposon Repression in Adult Ovaries. *Mol Cell* **66**: 411-419.e4.

568 Akulenko N, Ryazansky S, Morgunova V, Komarov PA, Olovnikov I, Vaury C, Jensen S,
569 Kalmykova A. 2018. Transcriptional and chromatin changes accompanying de novo
570 formation of transgenic piRNA clusters. *RNA* **24**: 574–584.

571 Aravin A a, Hannon GJ, Brennecke J. 2007. The Piwi-piRNA pathway provides an adaptive
572 defense in the transposon arms race. *Science* **318**: 761–4.
573 http://www.ncbi.nlm.nih.gov/pubmed/17975059.

574 Bartolomé C, Maside X, Charlesworth B. 2002a. On the Abundance and Distribution of
575 Transposable Elements in the Genome of Drosophila melanogaster. *Mol Biol Evol* **19**: 926–
576 937.

577 Bartolomé C, Maside X, Charlesworth B. 2002b. On the abundance and distribution of
578 transposable elements in the genome of Drosophila melanogaster. *Mol Biol Evol* **19**: 926–
579 37.

580 Baumgartner L, Handler D, Platzer SW, Yu C, Duchek P, Brennecke J. 2022. The Drosophila
581 ZAD zinc finger protein Kipferl guides Rhino to piRNA clusters. *Elife* **11**.

582 Bergman CM, Quesneville H, Anxolabéhère D, Ashburner M. 2006. Recurrent insertion and
583 duplication generate networks of transposable element sequences in the *Drosophila*
584 *melanogaster* genome. *Genome Biol* **7**: R112.

585 Bertocchi NA, Torres FP, Deprá M, da Silva Valente VL. 2020. Evolutionary study of the
586 *412/mdg1* lineage of the *Ty3/gypsy* group of LTR retrotransposons in Diptera. *bioRxiv*
587 2020.09.24.311225.

588 Bingham PM, Chapman CH. 1986. Evidence that white-blood is a novel type of temperature-
589 sensitive mutation resulting from temperature-dependent effects of a transposon insertion
590 on formation of white transcripts. *EMBO J* **5**: 3343–51.

591 Bingham PM, Kidwell MG, Rubin GM. 1982. The molecular basis of P-M hybrid dysgenesis: The
592 role of the P element, a P-strain-specific transposon family. *Cell* **29**: 995–1004.

593 Blumenstiel JP, Erwin AA, Hemmer LW. 2016. What Drives Positive Selection in the Drosophila
594 piRNA Machinery? The Genomic Autoimmunity Hypothesis. *Yale J Biol Med* **89**: 499–512.

595 Blumenstiel JP, Hartl DL, Lozovsky ER. 2002. Patterns of insertion and deletion in contrasting
596 chromatin domains. *Mol Biol Evol* **19**: 2211–25.

597 Bowen NJ, McDonald JF. 2001. Drosophila Euchromatic LTR Retrotransposons are Much
598 Younger Than the Host Species in Which They Reside. *Genome Res* **11**: 1527.

599 Brennecke J, Aravin A a, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ. 2007.
600 Discrete small RNA-generating loci as master regulators of transposon activity in
601 Drosophila. *Cell* **128**: 1089–103.

602 Brennecke J, Malone CD, Aravin AA, Sachidanandam R, Stark A, Hannon GJ. 2008. An
603 Epigenetic Role for Maternally Inherited piRNAs in Transposon Silencing. *Science (1979)*
604 **322**: 1387–1392.

605 Bucheton A, Paro R, Sang HM, Pelisson A, Finnegan DJ. 1984. The molecular basis of I-R
606 hybrid dysgenesis in Drosophila melanogaster: identification, cloning, and properties of the
607 I factor. *Cell* **38**: 153–63.

608 Carr M, Bensasson D, Bergman CM. 2012. Evolutionary Genomics of Transposable Elements
609 in Saccharomyces cerevisiae. *PLoS One* **7**: e50978.

610 Chakraborty M, Chang C-H, Khost DE, Vedanayagam J, Adrion JR, Liao Y, Montooth KL,
611 Meiklejohn CD, Larracuente AM, Emerson JJ. 2021. Evolution of genome structure in the
612 *Drosophila simulans* species complex. *Genome Res* **31**: 380–396.

613 Chakraborty M, Emerson JJ, Macdonald SJ, Long AD. 2019. Structural variants exhibit
614 widespread allelic heterogeneity and shape variation in complex traits. *Nat Commun* **10**: 1–
615 11.

Chang N-C, Rovira Q, Wells J, Feschotte C, Vaquerizas JM. 2022. Zebrafish transposable elements show extensive diversification in age, genomic distribution, and developmental expression. *Genome Res* **32**: 1408–1423. http://genome.cshlp.org/lookup/doi/10.1101/gr.275655.121 (Accessed August 18, 2021).

Charlesworth B, Langley CH. 1989. The population genetics of Drosophila transposable elements. *Annu Rev Genet* **23**: 251–287.

Chen P, Kotov AA, Godneeva BK, Bazylev SS, Olenina L V., Aravin AA. 2021. piRNA-mediated gene regulation and adaptation to sex-specific transposon expression in D. melanogaster male germline. *Genes Dev* **38**.

Chirn G-W, Rahman R, Sytnikova YA, Matts JA, Zeng M, Gerlach D, Yu M, Berger B, Naramura M, Kile BT, et al. 2015. Conserved piRNA Expression from a Distinct Set of piRNA Cluster Loci in Eutherian Mammals. *PLoS Genet* **11**: e1005652.

Cosby RL, Chang N-C, Feschotte C. 2019. Host–transposon interactions: conflict, cooperation, and cooption. *Genes Dev* **33**: 1098–1116.

Costas J, Valadé E, Naveira H. 2001. Structural Features of the mdg1 Lineage of the Ty3/gypsy Group of LTR Retrotransposons Inferred from the Phylogenetic Analyses of Its Open Reading Frames. *J Mol Evol* **53**: 165–171.

Di Nocera PP, Graziani F, Lavorgna G. 1986. Genomic and structural organization of Drosophila melanogaster G elements. *Nucleic Acids Res* **14**: 675–91.

Dolgin ES, Charlesworth B. 2006. The fate of transposable elements in asexual populations. *Genetics* **174**: 817–827.

Dopman EB, Hartl DL. 2007. A portrait of copy-number polymorphism in Drosophila melanogaster. *Proceedings of the National Academy of Sciences* **104**: 19920–19925.

Ellison CE, Cao W. 2020. Nanopore sequencing and Hi-C scaffolding provide insight into the evolutionary dynamics of transposable elements and piRNA production in wild strains of Drosophila melanogaster. *Nucleic Acids Res* **48**: 290–303.

Ewing AD, Smits N, Sanchez-Luque FJ, Faivre J, Brennan PM, Richardson SR, Cheetham SW, Faulkner GJ. 2020. Nanopore Sequencing Enables Comprehensive Transposable Element Epigenomic Profiling. *Mol Cell* **80**: 915-928.e5.

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* **117**: 9451–9457.

Gebert D, Neubert LK, Lloyd C, Gui J, Lehmann R, Teixeira FK. 2021. Large Drosophila germline piRNA clusters are evolutionarily labile and dispensable for transposon regulation. *Mol Cell* **81**: 3965-3978.e5.

Genzor P, Konstantinidou P, Stoyko D, Manzourolajdad A, Marlin Andrews C, Elchert AR, Stathopoulos C, Haase AD. 2021. Cellular abundance shapes function in piRNA-guided genome defense. *Genome Res* **31**: 2058–2068.

Grimson A, Srivastava M, Fahey B, Woodcroft BJ, Chiang HR, King N, Degnan BM, Rokhsar DS, Bartel DP. 2008. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature 2008 455:7217* **455**: 1193–1197.

Han S, Dias GB, Basting PJ, Viswanatha R, Perrimon N, Bergman CM. 2022. Local assembly of long reads enables phylogenomics of transposable elements in a polyploid cell line. *Nucleic Acids Res* **50**: e124.

Heller D, Vingron M. 2019. SVIM: structural variant identification using mapped long reads. *Bioinformatics* **35**: 2907–2915.

Houwing S, Kamminga LM, Berezikov E, Cronembold D, Girard A, van den Elst H, Filippov D V., Blaser H, Raz E, Moens CB, et al. 2007. A Role for Piwi and piRNAs in Germ Cell Maintenance and Transposon Silencing in Zebrafish. *Cell* **129**: 69–82.

Huang W, Massouras A, Inoue Y, Peiffer J, Ramia M, Tarone AM, Turlapati L, Zichner T, Zhu D, Lyman RF, et al. 2014. Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. *Genome Res* **24**: 1193–1208.

Huang Y, Shukla H, Lee YCG. 2022. Species-specific chromatin landscape determines how transposable elements shape genome evolution. *Elife* **11**.

Hung YH, Slotkin RK. 2021. The initiation of RNA interference (RNAi) in plants. *Curr Opin Plant Biol* **61**. https://pubmed.ncbi.nlm.nih.gov/33657510/ (Accessed July 25, 2022).

Inouye S, Hattori K, Yuki S, Saigo K. 1986. Structural variations in the *Drosophila* retrotransposon, *17.6*. *Nucleic Acids Res* **14**: 4765–4778.

Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, Liu Y, Liu B, Wang Y. 2020. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol* **21**: 1–24.

Kaminker JS, Bergman CM, Kronmiller B, Carlson J, Svirskas R, Patel S, Frise E, Wheeler DA, Lewis SE, Rubin GM, et al. 2002. The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. *Genome Biol* **3**: research0084.1.

Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay LA, Bourque G, Yandell M, Feschotte C. 2013. Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. *PLoS Genet* **9**: e1003470.

Kelleher ES, Azevedo RBR, Zheng Y. 2018. The Evolution of Small-RNA-Mediated Silencing of an Invading Transposable Element. *Genome Biol Evol* **10**: 3038–3057.

Kelleher ES, Barbash DA. 2013. Analysis of piRNA-mediated silencing of active TEs in Drosophila melanogaster suggests limits on the evolution of host genome defense. *Mol Biol Evol* **30**: 1816–29.

Khurana JS, Wang J, Xu J, Koppetsch BS, Thomson TC, Nowosielska A, Li C, Zamore PD, Weng Z, Theurkauf WE. 2011. Adaptation to P element transposon invasion in drosophila melanogaster. *Cell* **147**: 1551–1563.

Klattenhoff C, Xi H, Li C, Lee S, Xu J, Khurana JS, Zhang F, Schultz N, Koppetsch BS, Nowosielska A, et al. 2009. The Drosophila HP1 homolog Rhino is required for transposon silencing and piRNA production by dual-strand clusters. *Cell* **138**: 1137–49.

693 Kofler R. 2019. Dynamics of Transposable Element Invasions with piRNA Clusters. *Mol Biol*
694 *Evol* **36**: 1457–1472.

695 Kofler R. 2020. piRNA Clusters Need a Minimum Size to Control Transposable Element
696 Invasions. *Genome Biol Evol* **12**: 736–749.

697 Kofler R, Nolte V, Schlötterer C. 2015. Tempo and Mode of Transposable Element Activity in
698 Drosophila. *PLoS Genet* **11**: e1005406.

699 Larkin A, Marygold SJ, Antonazzo G, Attrill H, dos Santos G, Garapati P V, Goodman JL,
700 Gramates LS, Millburn G, Strelets VB, et al. 2021. FlyBase: updates to the Drosophila
701 melanogaster knowledge base. *Nucleic Acids Res* **49**: D899–D907.

702 Lau NC, Seto AG, Kim J, Kuramochi-Miyagawa S, Nakano T, Bartel DP, Kingston RE. 2006.
703 Characterization of the piRNA complex from rat testes. *Science* **313**: 363–7.

704 Lawlor MA, Cao W, Ellison CE. 2021. A transposon expression burst accompanies the
705 activation of Y-chromosome fertility genes during Drosophila spermatogenesis. *Nat*
706 *Commun* **12**: 6854.

707 Le Thomas A, Stuwe E, Li S, Du J, Marinov G, Rozhkov N, Chen Y-CA, Luo Y, Sachidanandam
708 R, Toth KF, et al. 2014. Transgenerationally inherited piRNAs trigger piRNA biogenesis by
709 changing the chromatin of piRNA clusters and inducing precursor processing. *Genes Dev*
710 **28**: 1667–80.

711 Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–
712 3100.

713 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R.
714 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

715 Ma Q, Srivastav SP, Gamez S, Dayama G, Feitosa-Suntheimer F, Patterson EI, Johnson RM,
716 Matson EM, Gold AS, Brackney DE, et al. 2021. A mosquito small RNA genomics resource
717 reveals dynamic evolution and host responses to viruses and transposons. *Genome Res*
718 **31**: 512–528.

719 Miller DE, Dorador AP, Van Vaerenberghe K, Li A, Grantham EK, Cerbin S, Cummings C,
720 Barragan M, Egidy RR, Scott AR, et al. 2023. Off-target piRNA gene silencing in
721 Drosophila melanogaster rescued by a transposable element insertion. *PLoS Genet* **19**:
722 e1010598.

723 Mohn F, Sienski G, Handler D, Brennecke J. 2014. The Rhino-Deadlock-Cutoff Complex
724 Licenses Noncanonical Transcription of Dual-Strand piRNA Clusters in Drosophila. *Cell*
725 **157**: 1364–1379.

726 Montgomery EA, Huang SM, Langley CH, Judd BH. 1991. Chromosome rearrangement by
727 ectopic recombination in Drosophila melanogaster: genome structure and evolution.
728 *Genetics* **129**: 1085–1098.

729 Olovnikov I, Ryazansky S, Shpiz S, Lavrov S, Abramov Y, Vaury C, Jensen S, Kalmykova A.
730 2013. De novo piRNA cluster formation in the Drosophila germ line triggered by transgenes
731 containing a transcribed transposon fragment. *Nucleic Acids Res* **41**: 5757–5768.

732 Ozata DM, Gainetdinov I, Zoch A, O'Carroll D, Zamore PD. 2019. PIWI-interacting RNAs: small
733 RNAs with big functions. *Nat Rev Genet* **20**: 89–108.

734 Palmer WH, Hadfield JD, Obbard DJ. 2018. RNA-Interference Pathways Display High Rates of
735 Adaptive Protein Evolution in Multiple Invertebrates. *Genetics* **208**: 1585–1599.

736 Pasquesi GIM, Perry BW, Vandewege MW, Ruggiero RP, Schield DR, Castoe TA. 2020.
737 Vertebrate Lineages Exhibit Diverse Patterns of Transposable Element Regulation and
738 Expression across Tissues. *Genome Biol Evol* **12**: 506–521.

739 Petrov D a., Aminetzach YT, Davis JC, Bensasson D, Hirsh AE. 2003. Size matters: Non-LTR
740 retrotransposable elements and ectopic recombination in drosophila. *Mol Biol Evol* **20**:
741 880–892.

742 Rech GE, Radío S, Guirao-Rico S, Aguilera L, Horvath V, Green L, Lindstadt H, Jamilloux V,
743 Quesneville H, González J. 2022. Population-scale long-read sequencing uncovers
744 transposable elements associated with gene expression variation and adaptive signatures
745 in Drosophila. *Nat Commun* **13**: 1948.

746 Riddle NC, Minoda A, Kharchenko P V., Alekseyenko AA, Schwartz YB, Tolstorukov MY,
747 Gorchakov AA, Jaffe JD, Kennedy C, Linder-Basso D, et al. 2011. Plasticity in patterns of
748 histone modifications and chromosomal proteins in Drosophila heterochromatin. *Genome*
749 *Res* **21**: 147–163.

750 Rosenkranz D, Zischler H. 2012. proTRAC--a software for probabilistic piRNA cluster detection,
751 visualization and analysis. *BMC Bioinformatics* **13**: 5.

752 Rozhkov N V, Aravin AA, Zelentsova ES, Schostak NG, Sachidanandam R, McCombie WR,
753 Hannon GJ, Evgen'ev MB. 2010. Small RNA-based silencing strategies for transposons in
754 the process of invading Drosophila species. *RNA* **16**: 1634–45.

755 Russo J, Harrington AW, Steiniger M. 2016. Antisense Transcription of Retrotransposons in
756 Drosophila: An Origin of Endogenous Small Interfering RNA Precursors. *Genetics* **202**:
757 107–21.

758 Said I, McGurk MP, Clark AG, Barbash DA. 2022. Patterns of piRNA Regulation in Drosophila
759 Revealed through Transposable Element Clade Inference. *Mol Biol Evol* **39**.

760 Saint-Leandre B, Capy P, Hua-Van A, Filee J. 2020. piRNA and Transposon Dynamics in
761 Drosophila: A Female Story. *Genome Biol Evol* **12**: 931–947.

762 Schrider DR, Houle D, Lynch M, Hahn MW. 2013. Rates and genomic consequences of
763 spontaneous mutational events in Drosophila melanogaster. *Genetics* **194**: 937–54.

764 Sentmanat MF, Elgin SCR. 2012. Ectopic assembly of heterochromatin in Drosophila
765 melanogaster triggered by transposable elements. *Proc Natl Acad Sci U S A* **109**: 14104–
766 14109.

767 Shpiz S, Ryazansky S, Olovnikov I, Abramov Y, Kalmykova A. 2014. Euchromatic Transposon
768 Insertions Trigger Production of Novel Pi- and Endo-siRNAs at the Target Sites in the
769 Drosophila Germline. *PLoS Genet* **10**: e1004138.

Simkin A, Wong A, Poh Y-P, Theurkauf WE, Jensen JD. 2013. Recurrent and recent selective sweeps in the piRNA pathway. *Evolution* **67**: 1081.

Smit AF. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev* **9**: 657–663.

Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, Kalef-Ezra E, Gandhi M, Hong K, Pehlivan D, et al. 2024. Detection of mosaic and population-level structural variants with Sniffles2. *Nat Biotechnol*.

Solares EA, Chakraborty M, Miller DE, Kalsow S, Hall K, Perera AG, Emerson JJ, Hawley RS. 2018. Rapid Low-Cost Assembly of the *Drosophila melanogaster* Reference Genome Using Low-Coverage, Long-Read Sequencing. *G3 Genes|Genomes|Genetics* **8**: 3143–3154.

Sultana T, Zamborlini A, Cristofari G, Lesage P. 2017. Integration site selection by retroviruses and transposable elements in eukaryotes. *Nature Reviews Genetics 2017 18:5* **18**: 292–308.

Watanabe T, Totoki Y, Toyoda A, Kaneda M, Kuramochi-Miyagawa S, Obata Y, Chiba H, Kohara Y, Kono T, Nakano T, et al. 2008. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature 2008 453:7194* **453**: 539–543.

Wells JN, Feschotte C. 2020. A Field Guide to Eukaryotic Transposable Elements. *Annu Rev Genet* **54**: 539.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, et al. 2007. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* **8**: 973–982.

Wierzbicki F, Kofler R, Signor S. 2023. Evolutionary dynamics of piRNA clusters in Drosophila. *Mol Ecol* **32**: 1306–1322.

Yi M, Chen F, Luo M, Cheng Y, Zhao H, Cheng H, Zhou R. 2014. Rapid Evolution of piRNA Pathway in the Teleost Fish: Implication for an Adaptation to Transposon Diversity. *Genome Biol Evol* **6**: 1393–1407.

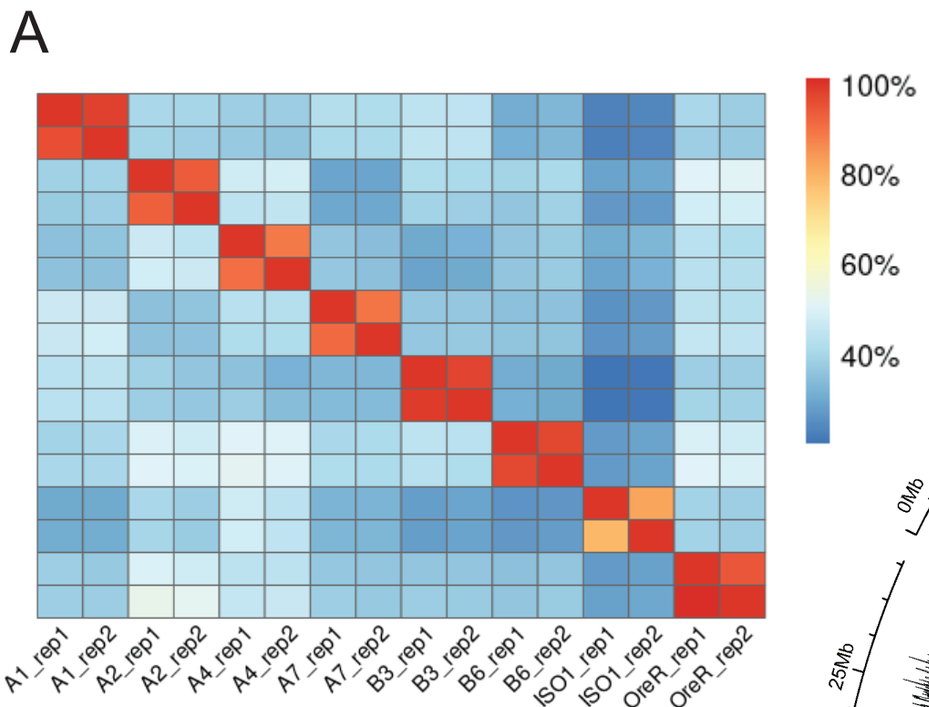
Yu T, Koppetsch BS, Pagliarani S, Johnston S, Silverstein NJ, Luban J, Chappell K, Weng Z, Theurkauf WE. 2019. The piRNA Response to Retroviral Invasion of the Koala Genome. *Cell* **179**: 632-643.e12.

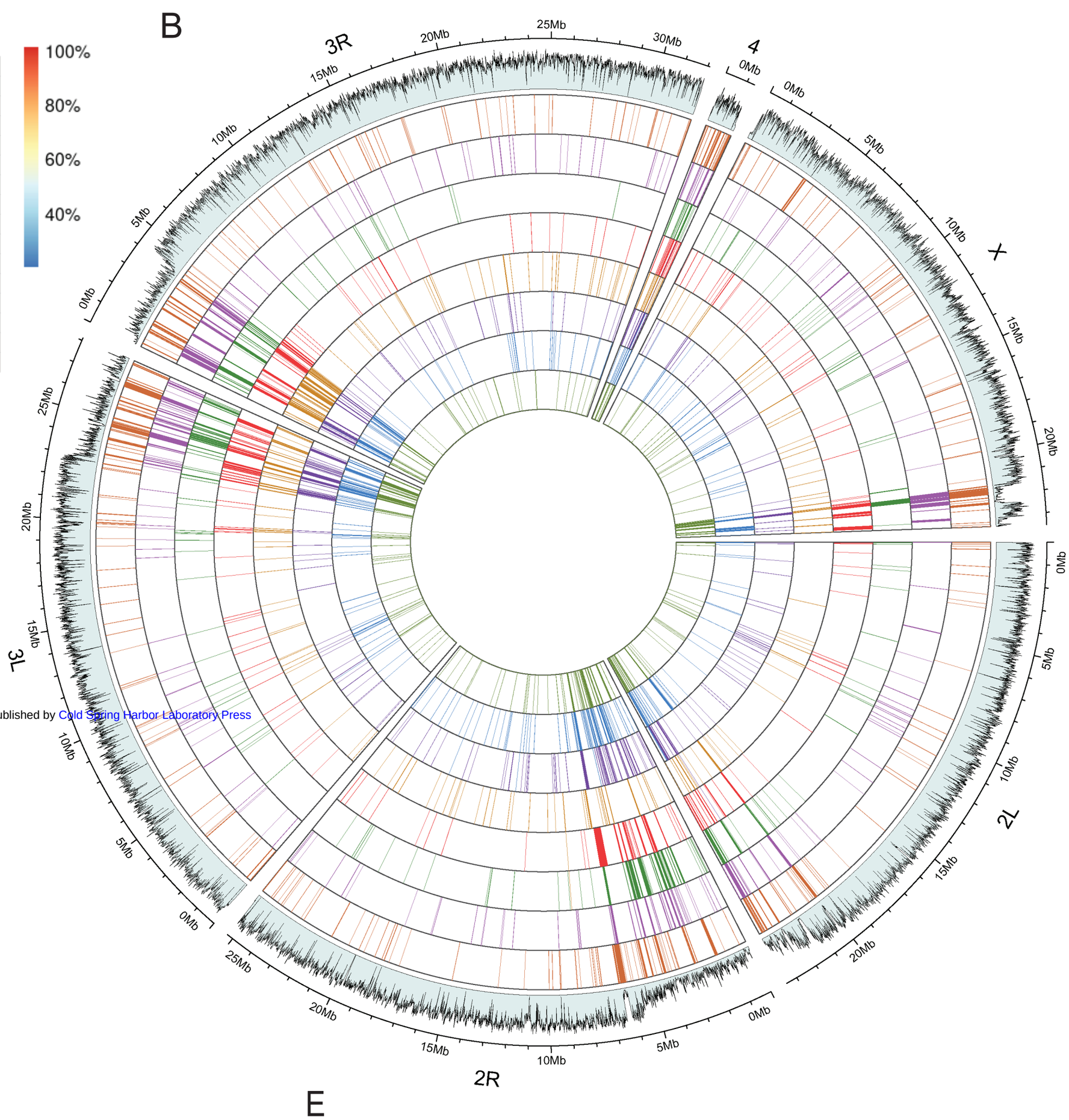
Zanni V, Eymery A, Coiffet M, Zytnicki M, Luyten I, Quesneville H, Vaury C, Jensen S. 2013. Distribution, evolution, and diversity of retrotransposons at the flamenco locus reflect the regulatory properties of piRNA clusters. *Proc Natl Acad Sci U S A* **110**: 19842–7.

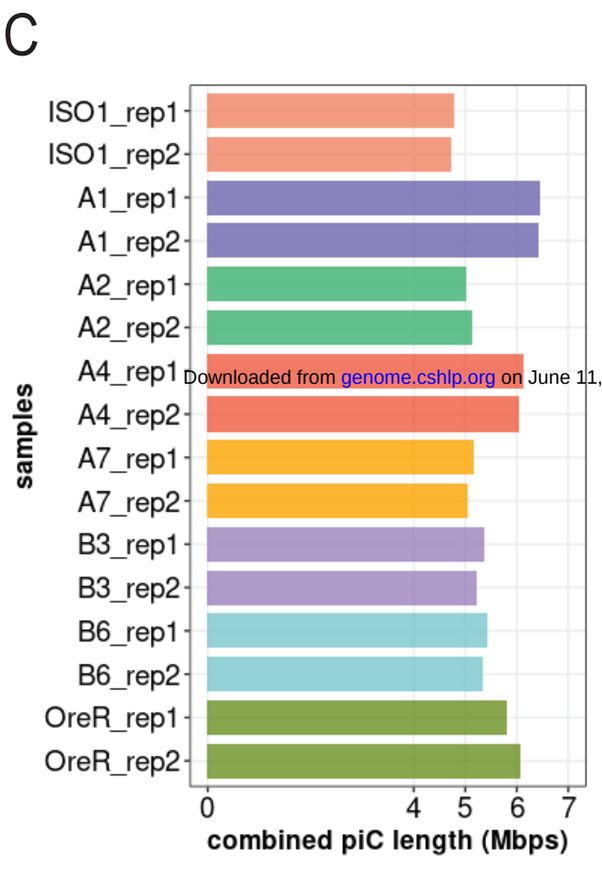
Zhang S, Pointer B, Kelleher ES. 2020. Rapid evolution of piRNA-mediated silencing of an invading transposable element was driven by abundant de novo mutations. *Genome Res* **30**: 566–575.

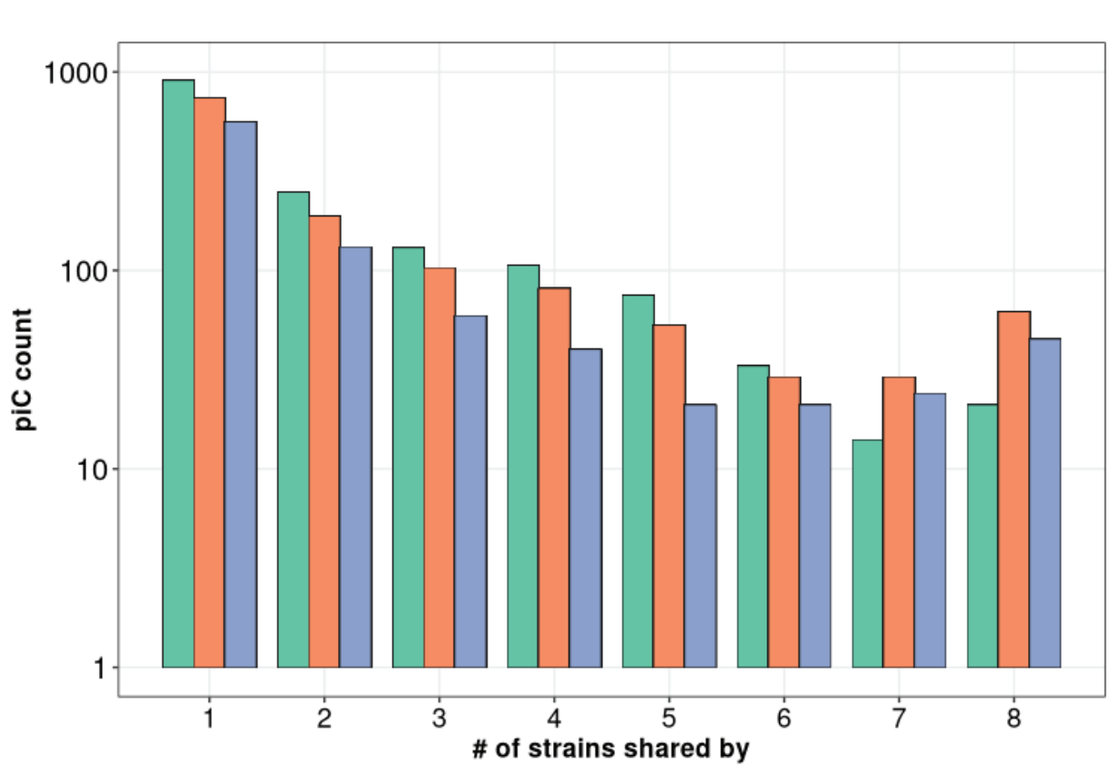
Zichner T, Garfield DA, Rausch T, Stütz AM, Cannavó E, Braun M, Furlong EEM, Korbel JO. 2013. Impact of genomic structural variation in *Drosophila melanogaster* based on population-scale sequencing. *Genome Res* **23**: 568–579.

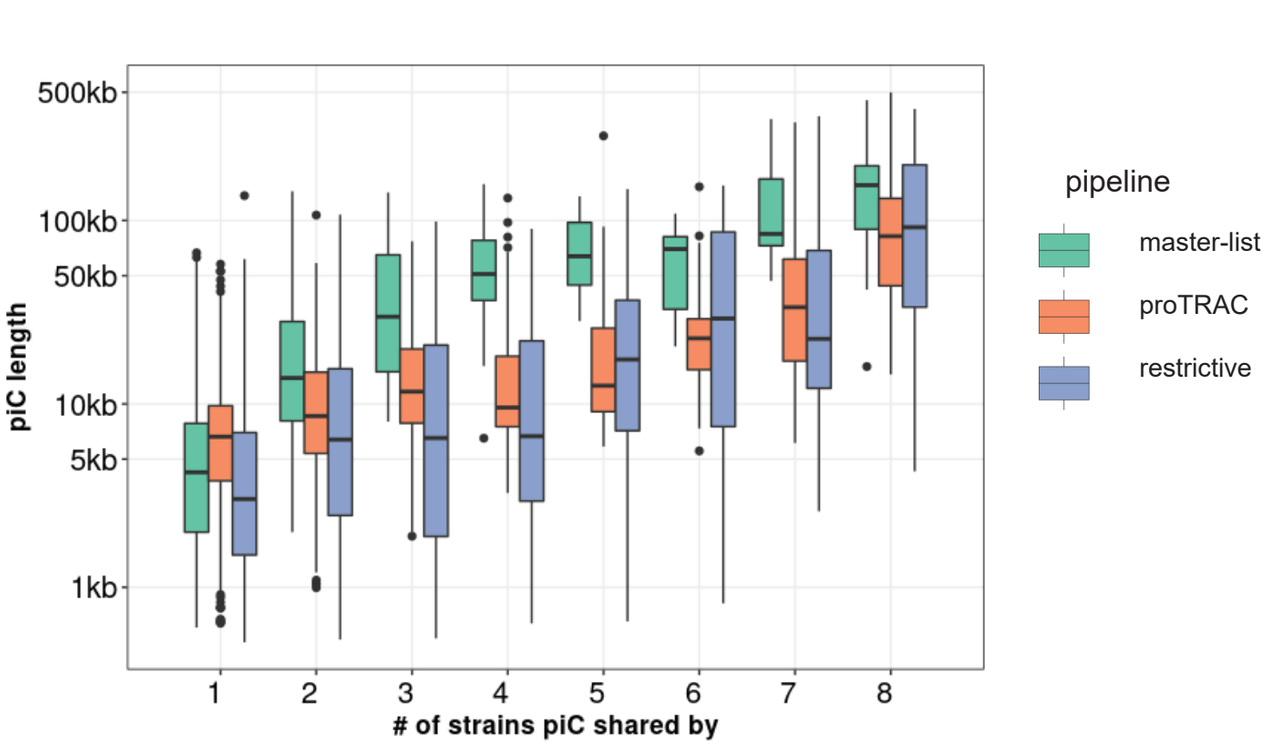
Figure 1



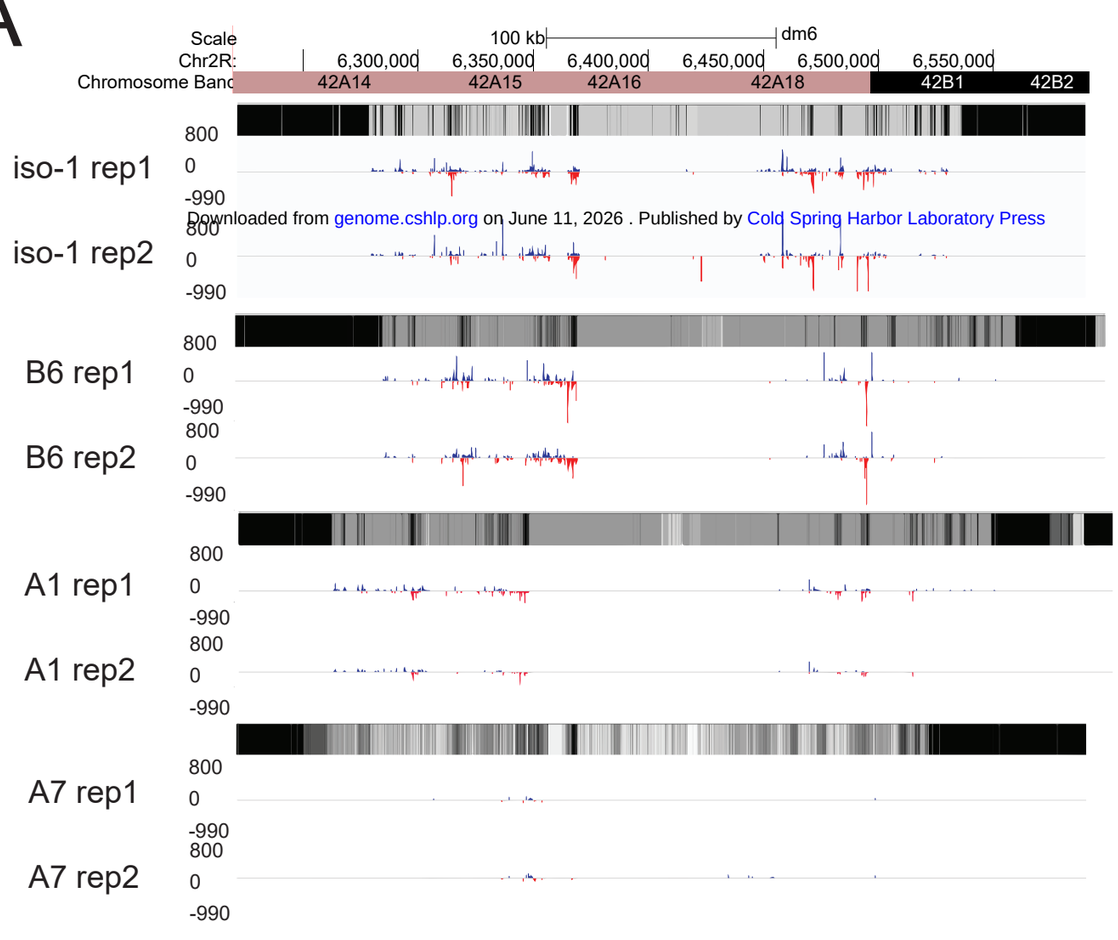




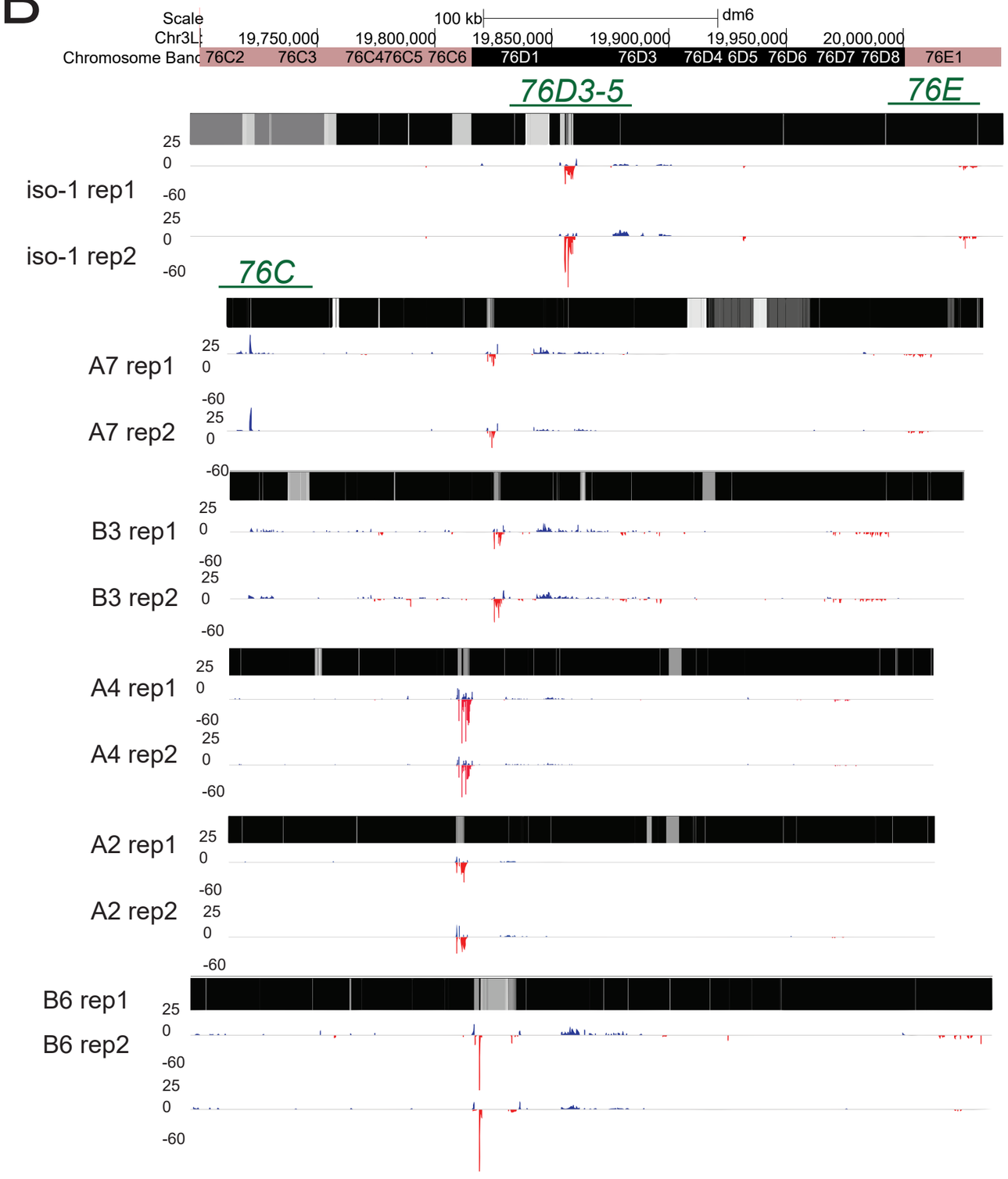




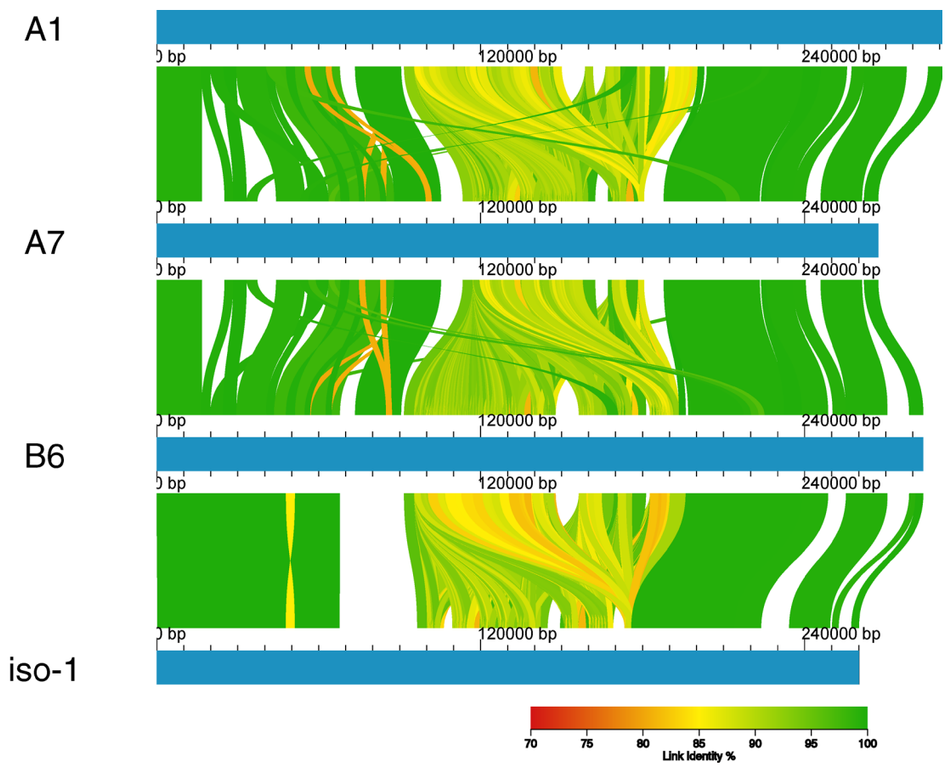
A



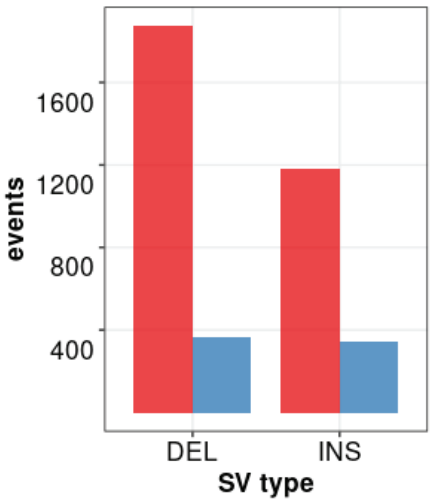
B



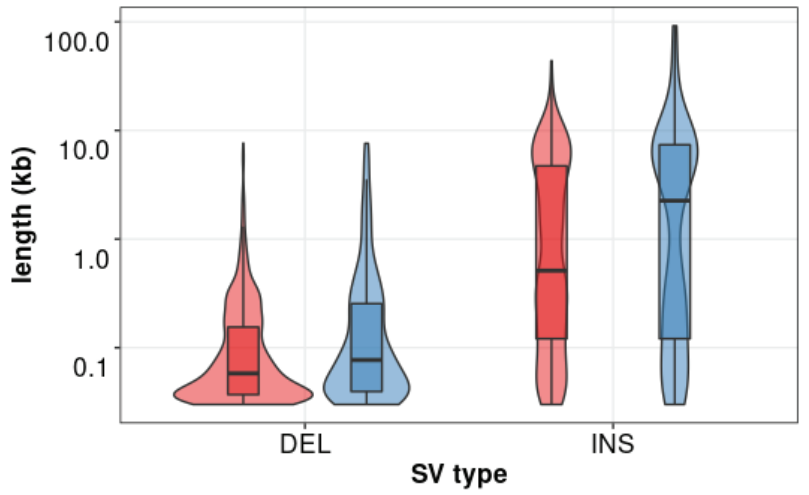
C



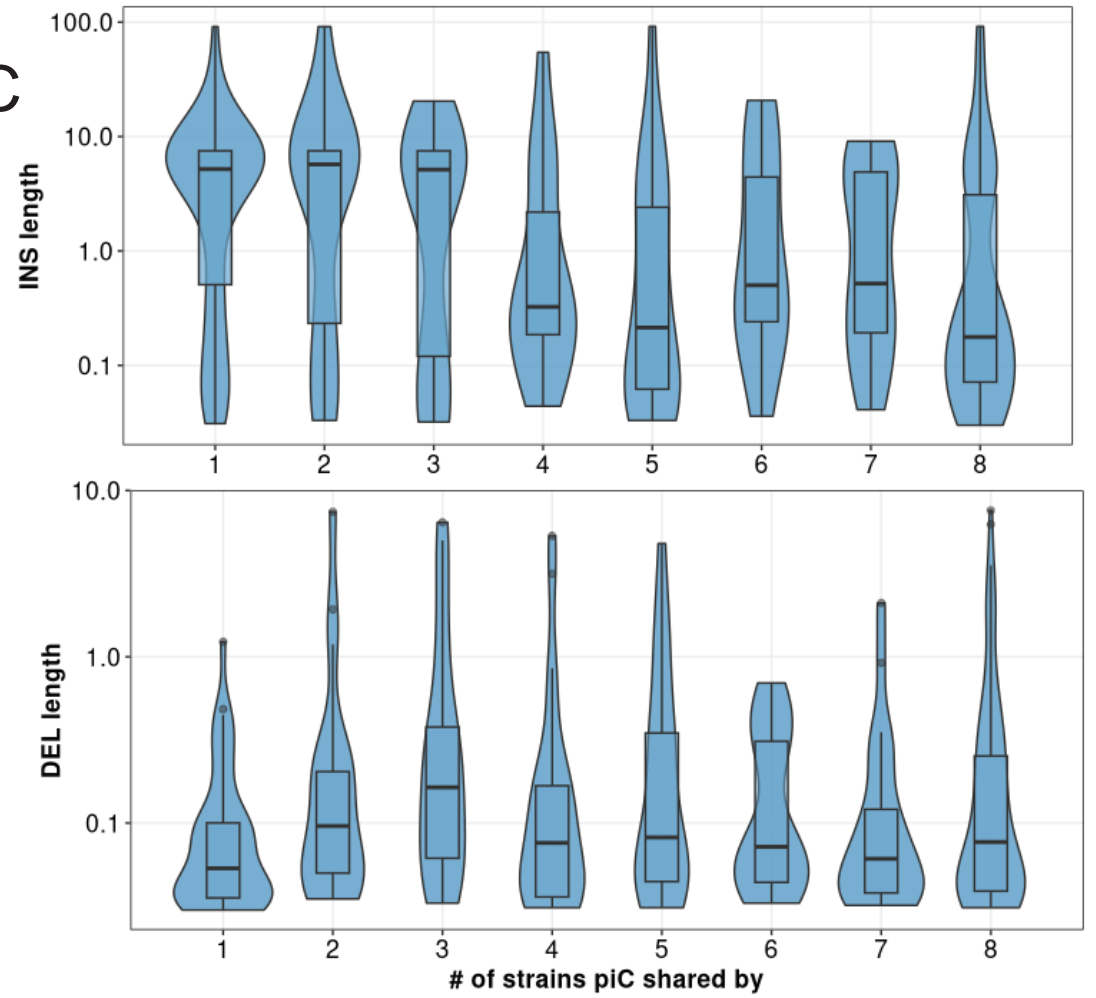
A



B

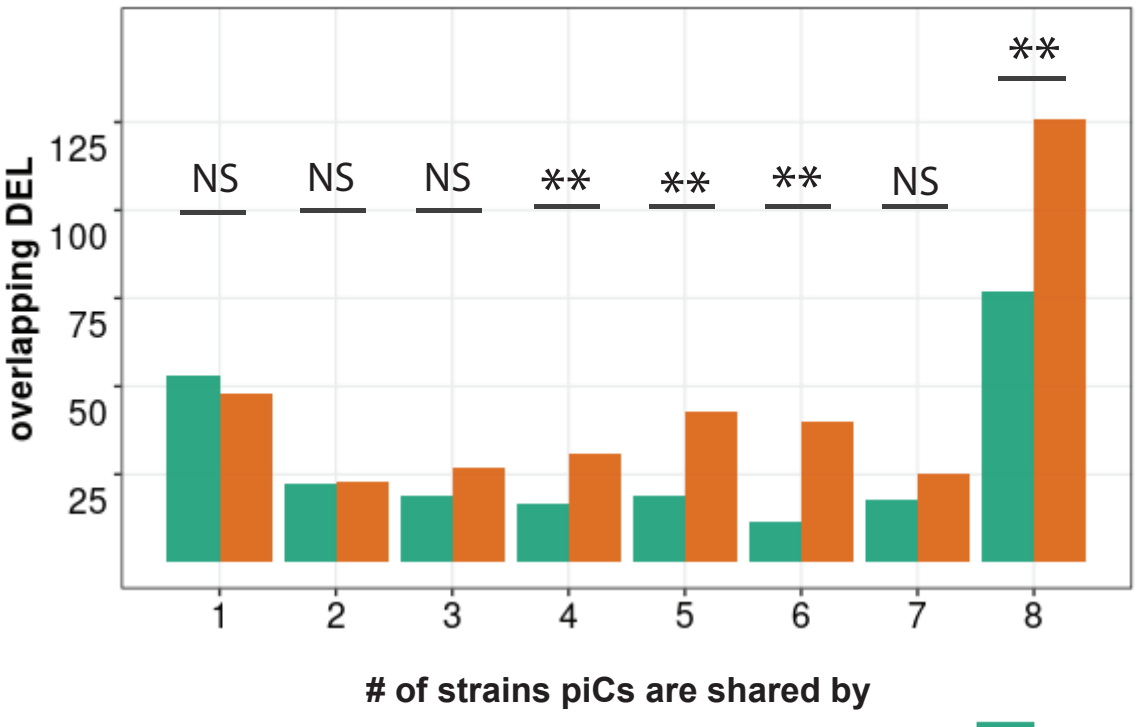


C

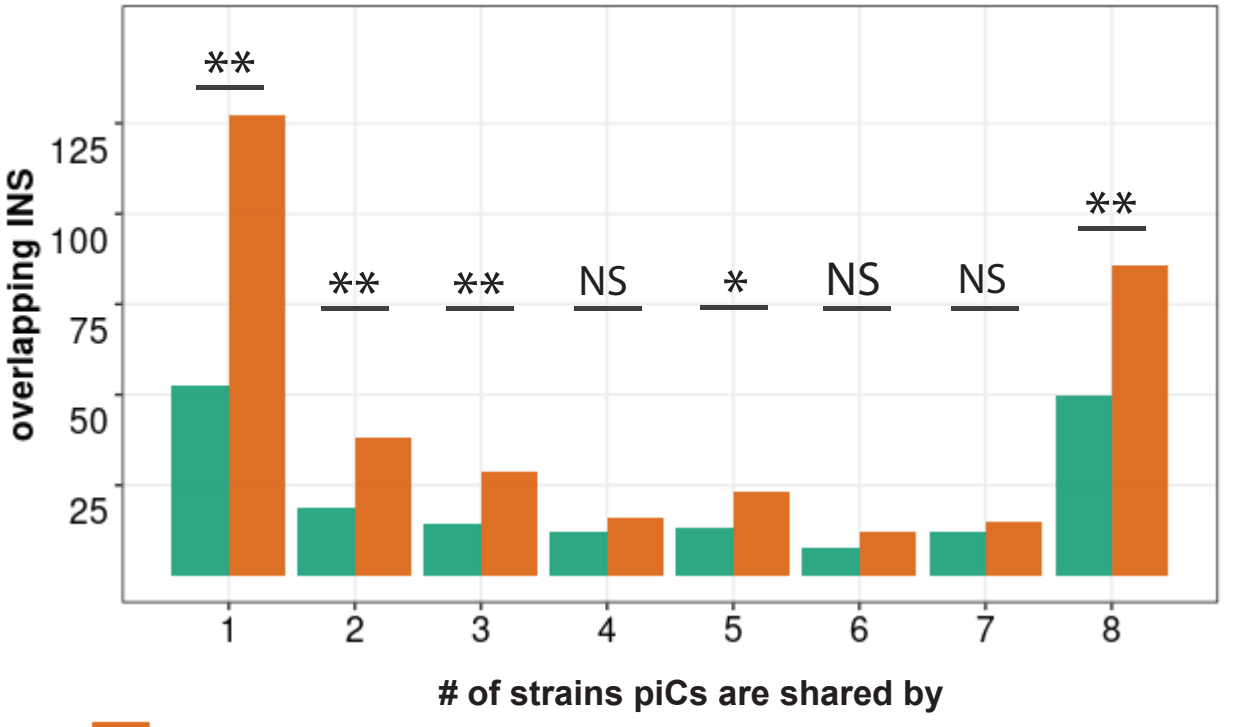


genome-wide
piC overlapping

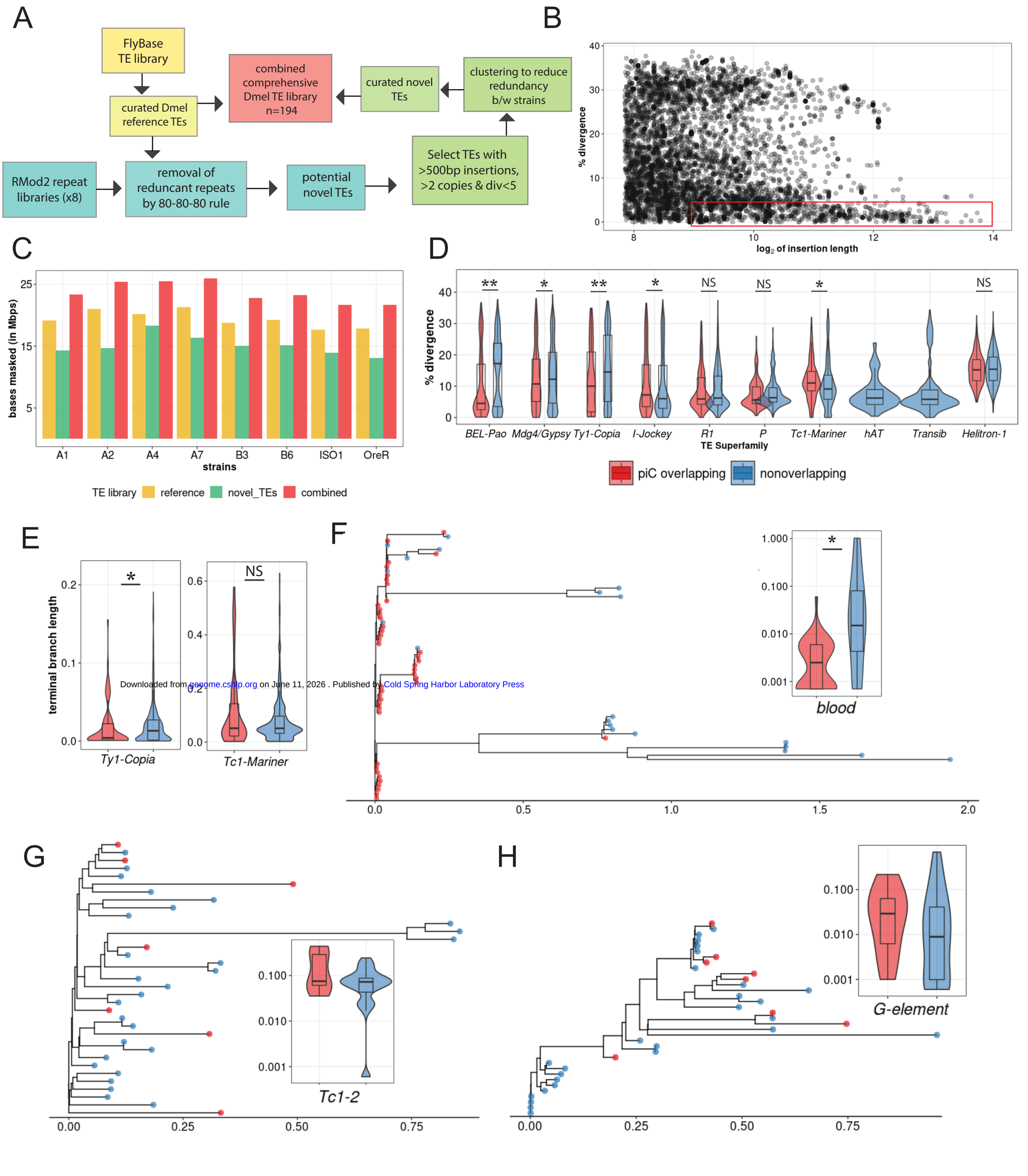
D



E



mean expected
observed



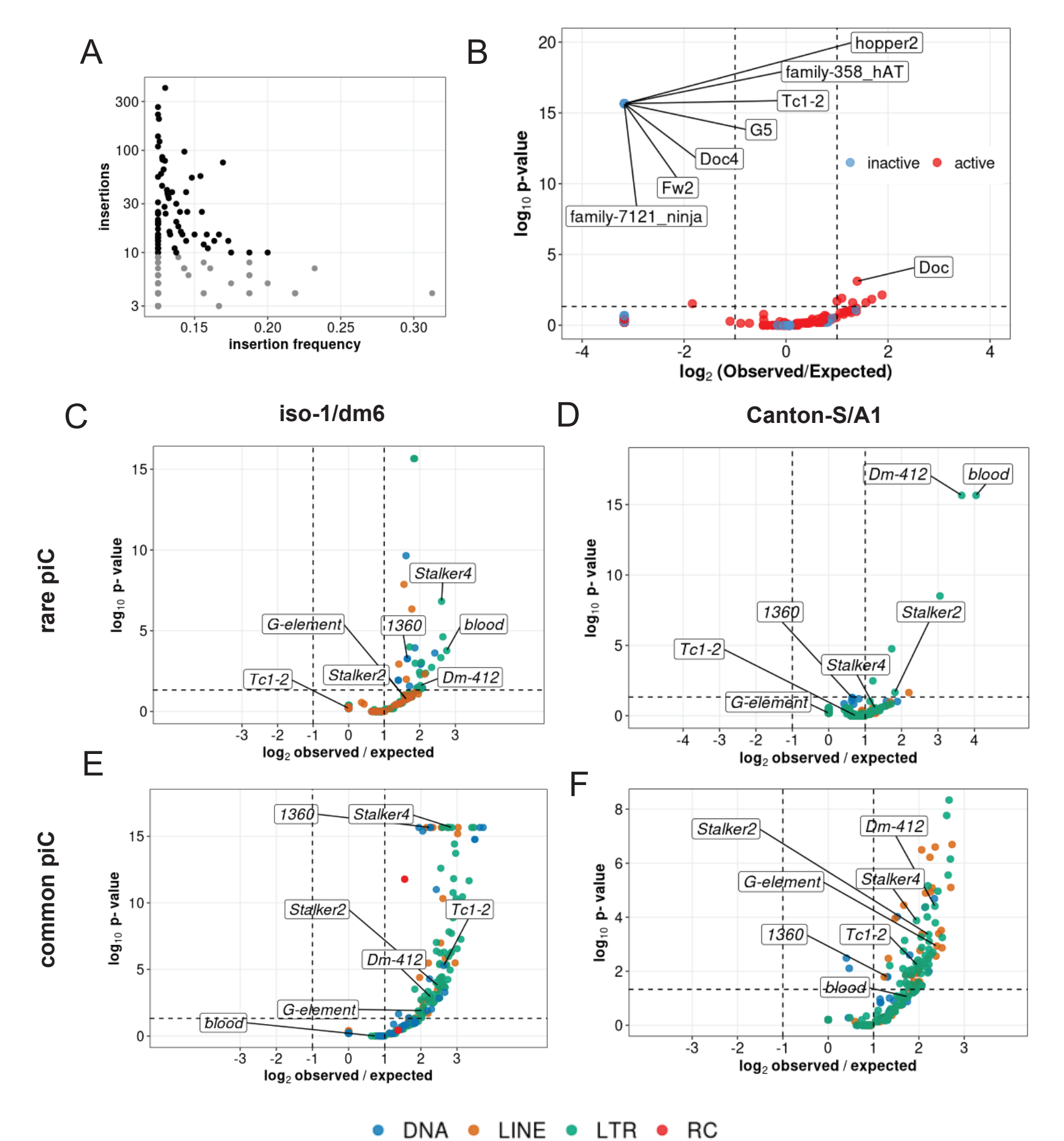

A
insertions
insertion frequency
B
hopper2
family-358_hAT
Tc1-2
G5
Doc4
Fw2
family-7121_ninja
Doc
inactive
active
log10 p-value
log2 (Observed/Expected)
C
iso-1/dm6
rare piC
Stalker4
G-element
1360
blood
Tc1-2
Stalker2
Dm-412
log10 p- value
log2 observed / expected
D
Canton-S/A1
Dm-412
blood
1360
Stalker2
Tc1-2
Stalker4
G-element
E
common piC
F
DNA
LINE
LTR
RC

Figure 6